EOSC POLICY BRIEF

CALL: HORIZON-WIDERA-2021-ERA-01-41
TOPIC: GLOBAL COOPERATION ON FAIR DATA POLICY AND PRACTICE
PROJECT: GLOBAL COOPERATION ON FAIR DATA POLICY AND PRACTICE (*WORLDFAIR*) HTTPS://WORLDFAIR-PROJECT.EU/
DATE: 05 JUNE 2024. DOI: 10.5281/ZENODO.11242702.

SCOPE OF THE POLICY BRIEF

In this policy brief, the projects contributing to the advancement of the European Open Science Cloud – EOSC- report on the progress made and provide recommendations to the European Commission for further policy analysis and development. This policy brief should be understood as complementary to the other mandatory reporting materials.

***Policy background**:*

The European Open Science Cloud (EOSC) is recognised by the Council of the European Union as the pilot action to deepen the new European Research Area (ERA) and is included in the ERA policy agenda 2022-2024. EOSC is also recognised in the European strategy for data as the data space for science, research and innovation which shall be fully articulated with the other sectoral data spaces defined in the strategy.

Overall progress is steered by the EOSC tripartite governance involving the Union represented by the European Commission, the participating countries represented in the EOSC Steering Board and the research community represented by the EOSC Association. The second phase of development of EOSC (2021-2030) takes place in the context of the EOSC European co-programmed Partnership, which brings together the European Commission and the EOSC Association according to the Strategic Research and Innovation Agenda (SRIA) which is co-developed with the entire EOSC community and sets three general objectives:

1. *Open science becomes the 'new normal', by ensuring that open science practices and skills are rewarded and taught.*
2. *Researchers can seamlessly find, access, reuse and combine results, through the definition of common standards and the development of related tools and services.*
3. *A federated infrastructure under community governance enabling open sharing of scientific results is deployed and sustained.*

FEEDBACK ON PROGRESS

1: Contribution to the SRIA: 'WorldFAIR: Global cooperation on FAIR data policy and practice'[1] was a two-year project to advance implementation of the FAIR principles[2], particularly in relation to interoperability and reusability of data within and across research domains. With an explicit mission to advance global collaboration and include partners from outside the European Union, WorldFAIR was coordinated by CODATA[3], the Committee on Data of the International Science Council[4] and with the Research Data Alliance (RDA) Association[5] as a major partner.

The project was conceived as responding to Recommendation 4 of the *Turning FAIR into Reality* report, which identified the need to 'Develop interoperability frameworks for FAIR sharing within disciplines and for interdisciplinary research'. The Recommendation states that 'Research communities need to be supported to develop interoperability frameworks that define their practices for data sharing, data formats, metadata standards, tools and infrastructure. To support interdisciplinary research, these interoperability frameworks should be articulated in common ways and adopt global standards where relevant.'[6] It is this vision of enabling community agreements around interoperability, and encouraging the identification and adoption of common standards, that drove the WorldFAIR project. This recommendation is cited in the SRIA[7] (pp.55-6) and is fundamental to achieving the vision of EOSC and Open Science more generally.

WorldFAIR worked with a set of eleven domain and cross-domain Case Studies[8], carefully chosen from existing CODATA and RDA activities to provide maximum impact. Each Case Study (in project terms, a 'Work Package') developed an interoperability framework, comprising policy and technical recommendations, for their discipline or interdisciplinary research area. Led by CODATA, a coordinating and synthesis activity (Work Package 2) supported each Case Study in understanding their requirements through the completion of FAIR Implementation Profiles (FIPs)[9]. In turn, these insights were incorporated into the development of a Cross-Domain Interoperability Framework (CDIF)[10] and recommendations for future work on domain-sensitive FAIR guidelines and assessment[11].

Each of these outputs — discipline-oriented interoperability frameworks and/or recommendations; a set of FIPs; a Cross-Domain Interoperability Framework and recommendations for future work on domain-sensitive FAIR assessment — is relevant to the SRIA, and those parts of the SRIA that deal with the EOSC Interoperability Framework[12], metadata and ontologies, and FAIR metrics and certification.

WorldFAIR is also relevant to the parts of the SRIA that deal with 'facing global challenges through multi-disciplinary programmes', the 'diversity of FAIR practices', and the need for 'community standards'. In sum, the WorldFAIR project makes a significant contribution to the EOSC Interoperability Framework in two important respects. First, the project underlines, and puts into practice, the need to engage research disciplines, ***at a global scale***, in the development of agreements and frameworks for FAIR. Second, the CDIF identifies a set of functional requirements to support FAIR and identifies existing or emerging

[1] https://worldfair-project.eu/
[2] Wilkinson, M. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
[3] https://codata.org/
[4] https://council.science/
[5] https://www.rd-alliance.org/rda-europe
[6] European Commission, Directorate-General for Research and Innovation, 'Turning FAIR into reality: final report and action plan from the European Commission expert group on FAIR data', Publications Office, 2018, https://data.europa.eu/doi/10.2777/1524; p.29.
[7] European Commission, Directorate-General for Research and Innovation, *Strategic Research and Innovation Agenda (SRIA) of the European Open Science Cloud (EOSC)*, v1.0, June 2021. Publications Office of the European Union, 2022, https://data.europa.eu/doi/10.2777/935288.
[8] https://worldfair-project.eu/case-studies-of-worldfair/
[9] https://www.go-fair.org/how-to-go-fair/fair-implementation-profile/
[10] https://worldfair-project.eu/cross-domain-interoperability-framework/
[11] https://worldfair-project.eu/fair-assessment/
[12] European Commission, Directorate-General for Research and Innovation, Corcho, O., et al., 'EOSC interoperability framework: report from the EOSC Executive Board Working Groups FAIR and Architecture', Publications Office, 2021, https://data.europa.eu/doi/10.2777/620649

standards and protocols that can be used across domains. The CDIF is categorically not a new standard intended to replace others, but a framework of existing cross-domain standards or emerging protocols, which will add significant detail and actionability to the EOSC Interoperability Framework.

In relation to the updated SRIA (Dec 2023) and multi-annual roadmap (MAR), WorldFAIR is mentioned as one of two projects working on "domain standards and developing models for cross-domain interoperability"[13]. It is suggested that, within the forthcoming work programme, "emphasis should shift to a maturity model for cross-domain interoperability, enabling and encouraging communities to increase the potential for reuse of data." WorldFAIR has shown how this can be done and a further project with similar methodologies would be a good way to address these objectives. The WorldFAIR synthesis reports argue that any maturity model and assessment of FAIR practices needs to be based on detailed insight into the discipline's approach and requirements; and that the best tool we currently have for this is the FAIR Implementation Profile[14]. We also argue that the CDIF has shown how a framework for cross-domain FAIR should be constructed[15]. Further work is required, but the CDIF should be the basis for any maturity model for cross-domain interoperability. Finally, the WorldFAIR case study approach is an effective way to mobilise communities to increase FAIRness and reusability of data. As argued in our policy recommendations below, the role of international representative and authoritative organisations should not be neglected.

Considering the MAR for 2025-27, the WorldFAIR partners can contribute to the areas relating to interoperability and reusability of research data (particularly for high priority, cross-domain research areas); the development, sustainability and governance of metadata schemas and other semantic artefacts; and the empowerment of research communities. As argued below in our policy recommendations, we urge EOSC, its partners and stakeholders, to move beyond a merely bibliographic approach to data stewardship and to enable an engineering approach, as a necessary precondition to building the web of FAIR data and services. Further development of the CDIF will be an essential component for cross-domain research, to improve legal and organisational interoperability, and to address the emerging challenges of AI. We also recommend maintaining the international dimension leveraged in WorldFAIR, and empowering organisations that represent research communities and develop and maintain metadata schema.

2: Interactions and Synergies with EOSC Stakeholders: WorldFAIR participated in the EOSC coordination meetings (September 2022, June 2023). In January 2023, meetings were arranged with colleagues at DG RTD and with the EOSC Association. WorldFAIR will also be represented, after the official end of the project, at the EOSC coordination meeting in June 2024.

WorldFAIR has also participated in numerous meetings arranged by or through the EOSC Association on technical infrastructures and data spaces. A session on WorldFAIR, and specifically the use of FIPs, was organised at the EOSC Symposium, Prague, November 2022[16]. Similarly, the CDIF was presented in a session at the EOSC Symposium, Madrid, September 2023[17]. WorldFAIR contributed actively to the EOSC Winter School, and in particularly to Opportunity Area 2[18]: CODATA will contribute to the ongoing work of this OA, in particular in relation to the CDIF.

The RDA Association was a major partner in the WorldFAIR project, leading the communications Work Package: thus, the activities of WorldFAIR and its Case Studies are being given visibility through sessions at the RDA plenary and webinar series. A well-attended event on the WorldFAIR project's Cross-Domain Interoperability Framework was held on 20 March 2023, co-located with the RDA Plenary in Gothenburg. The WorldFAIR project was extremely prominent at International Data Week 2023, held in Salzburg, 23-27

[13] Strategic Research and Innovation Agenda (SRIA) of the European Open Science Cloud (EOSC), v1.2, p.175: https://eosc.eu/wp-content/uploads/2023/12/20231114_SRIA_1.2_final2.pdf
[14] Hodson, S. (2024). WorldFAIR (D2.2) 'WorldFAIR's Experience with FIPs' https://doi.org/10.5281/zenodo.11236094; Gregory, A., & Hodson, S. (2024). WorldFAIR (D2.4) 'Recommendations and framework for FAIR Assessment within (and across) disciplines' https://doi.org/10.5281/zenodo.11242737
[15] Gregory, A., et al. (2024). WorldFAIR (D2.3) 'Cross-Domain Interoperability Framework (CDIF)' https://doi.org/10.5281/zenodo.11236871
[16] https://events.eoscfuture.eu/symposium2022/programme
[17] https://symposium23.eoscfuture.eu/programme/
[18] https://eosc.eu/oa2-metadata-ontologies-interoperability/

October as 'a Festival of Data'[19]. This was an extremely significant and valuable opportunity to showcase WorldFAIR's work and solicit feedback.

Simon Hodson (CODATA, project coordinator) and Hilde Orten (Sikt, WP06 Social Surveys) were members of the EOSC-A Semantic Interoperability Task Force[20] and efforts were made to engage the leadership of the TF (Wolmar Nyberg Åkerström). A presentation on WorldFAIR was given to the SI TF on 3 May 2023. Wolmar participated in the CODATA-DDI Dagstuhl Workshop in October 2023 that involved a number of the WorldFAIR case studies and work on CDIF modules[21]. The workshop also included discussions of the new RDA activities on FAIR mappings, which are reflected in the SI TF's recommendations[22] and relate to work in FAIR-Impact.

CODATA and the RDA Association are partners in the FAIR-Impact project; discussions have been held with the technical leadership of EOSC-Core; CODATA is also a key partner in the RDA TIGER project, led by RDA Association. All these links have been used, and will continue to be used, to encourage synergies where possible.

A number of the WorldFAIR Case Studies have connections with significant Horizon Europe Projects: WP04 Nanomaterials partners were active in NanoCommons[23] and are now prominent in PARC[24]; WP09 Biodiversity lead partner GBIF was involved in DISSCO[25]; WP11 Ocean Sciences lead was also a partner in Blue Cloud[26]; WP13 Cultural Heritage lead DRI has links to Europeana[27].

Particularly relevant to the work of WorldFAIR, WP06 Social Surveys partner Sikt led an EOSC Future Science Project 'Climate Neutral and Smart Cities', which used CDIF components for cross-domain data integration combining social surveys data with contextual environmental data.[28]

3: Challenges: Funded through the WIDERA call channel, it may be that the global and international character of WorldFAIR leads to it being viewed as something of an outlier in relation to the INFRA-EOSC projects. Nevertheless, WorldFAIR continued to make the case for the importance of the EOSC engaging with international developments, organisations and stakeholders, particularly in the development of community practices, norms and standards.

CODATA and other WorldFAIR partners will continue to engage as much as possible with the EOSC Association and INFRA-EOSC projects and make the case for the domain recommendations and Cross-Domain Interoperability Framework that emerges. As described in our sustainability plan, CODATA will explore and refine the approach in a set of coordinated activities as WorldFAIR+. In particular, some new Case Studies will be launched in 2024 and the CDIF Working Group and Advisory Group will be maintained and will continue their work.

4: Contribution of WorldFAIR: The WP02 coordination and synthesis activity, led by CODATA, has drawn attention to the utility of FIPs. The early report, D1.2 'FAIR Implementation Profiles (FIPs) in WorldFAIR: what have we learnt?'[29] has been downloaded over 1200 times[30]. The report D2.2[31] provides a follow-up summary of our experiences alongside recommendations to improve the FIPs approach.

[19] https://worldfair-project.eu/2023/11/10/idw2023-a-festival-of-data-with-the-worldfair-project/
[20] See https://www.eosc.eu/advisory-groups/semantic-interoperability
[21] https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/2946531350/2023+Defining+a+Core+Metadata+Framework+for+Cross-Domain+Data+Sharing+and+Reuse
[22] Nyberg Åkerström, W., et al. (2024). Developing and implementing the semantic interoperability recommendations of the EOSC Interoperability Framework (27 March 2024). EOSC Association AISBL. https://doi.org/10.5281/zenodo.10843882
[23] See https://www.nanocommons.eu/
[24] See https://worldfair-project.eu/2023/04/12/next-up-in-the-parc-fair-data-and-tools-webinar-series/
[25] See https://www.dissco.eu/
[26] See https://blue-cloud.org/
[27] See https://pro.europeana.eu/organisation/digital-repository-of-ireland
[28] https://eoscfuture.eu/data/climate-neutral-and-smart-cities/
[29] See https://worldfair-project.eu/2023/03/22/the-worldfair-projects-cross-domain-interoperability-framework-2/
[30] Gregory, Arofan, & Hodson, Simon. (2022). WorldFAIR Project (D2.1) 'FAIR Implementation Profiles (FIPs) in WorldFAIR: What Have We Learnt?', https://doi.org/10.5281/zenodo.7378109
[31] Hodson, S. (2024)., https://doi.org/10.5281/zenodo.11236094

The major contribution of the coordination and synthesis activity is the Cross-Domain Interoperability Framework (CDIF). This has been discussed and publicised through numerous events including webinars, sessions at the RDA Plenary and International Data Week. The draft discovery module[32], which was shared for public feedback, has been downloaded over 1000 times to date.

The WorldFAIR D14.3 Outreach and Engagement Report[33] provides a full summary of WorldFAIR's participation and amplification through events (pp.31-38), the WorldFAIR webinar series (pp. 28-31), and by other means.

Each WorldFAIR Case Study produced an interoperability framework comprising recommendations and examples for their discipline or interdisciplinary research area. All the WorldFAIR Case Studies produced one or more FIPs as examples. Many produced guidelines and training materials. All the substantive deliverables from WorldFAIR are listed in an appendix to this document: one can see immediately the wealth of useful material (guidelines, recommendations, etc.) for the research areas involved.

In sum, the WorldFAIR project will make a major contribution to FAIR policy and practice in eleven specific research areas, while also proposing the Cross-Domain Interoperability Framework which will assist interdisciplinary research fields to make their data more FAIR.

5: WorldFAIR Contribution to other EU Policy Priorities:

☒ Horizon Europe Missions Cancer, Climate Change, Oceans, Climate neutral and Smart cities, Soil-deal for Europe

To a significant degree, the vision for WorldFAIR builds on CODATA's work for the International Science Council's Action Plan Project 2.1 'Making Data Work for Global Grand Challenges'. The Cross-Domain Interoperability Framework will be significant for increasing the FAIRness and utility of data for each of the priority areas listed, which are by their nature interdisciplinary.

The work of WorldFAIR WPs 7 (Population Health), 8 (Urban Health), 11 (Ocean Science), and 12 (Disaster Risk Reduction) will have relevance for the Horizon Europe Missions 'Climate Neutral and Smart Cities', 'Oceans', and 'Climate Change' respectively.

[32] Cross-Domain Interoperability Framework (CDIF) Working Group, Richard, S., et al. (2023). Cross Domain Interoperability Framework (CDIF): Discovery Module (v01 draft for public consultation) (Version 01). Zenodo. https://doi.org/10.5281/zenodo.10252564

[33] Delipalta, A. (2024). WorldFAIR (D14.3) WorldFAIR Outreach and Engagement Report (Version 1). Zenodo. https://doi.org/10.5281/zenodo.11205263

WORLDFAIR

Introduction to the Policy Brief and Recommendations
This document presents the final WorldFAIR policy recommendations. Building on the First WorldFAIR Policy Brief[34], it provides a synthesis of the most important recommendations from the project in relation to key WorldFAIR outputs, as well as those directed towards specific stakeholders and for the European Open Science Cloud.

WorldFAIR: Global cooperation on FAIR data policy and practice
'WorldFAIR: Global cooperation on FAIR data policy and practice'[35] was a two-year project to advance implementation of the FAIR principles[36], particularly in relation to interoperability and reusability of data, within and across research domains. The project had genuinely global scope, and an explicit mission to advance global collaboration and included partners from outside the European Union[37]. WorldFAIR was coordinated by CODATA[38], the Committee on Data of the International Science Council[39], with the Research Data Alliance (RDA) association[40] as a major partner. As well as these two international data organisations, WorldFAIR involved (either directly as beneficiaries, or indirectly through partner organisations) a number of authoritative and representative international projects, organisations, initiatives or infrastructures (e.g. NanoCommons[41], OneGeochemistry[42], SALURBAL[43], GBIF[44], ODIS[45]), as well as standards-setting organisations with global scope (e.g. IUPAC[46], DDI Alliance[47], OHDSI[48], TDWG[49]).

The project was conceived as responding to Recommendation 4 of the 'Turning FAIR into Reality' report, which identified the need to 'Develop interoperability frameworks for FAIR sharing within disciplines and for interdisciplinary research'. The Recommendation states that 'Research communities need to be supported to develop interoperability frameworks that define their practices for data sharing, data formats, metadata standards, tools and infrastructure. To support interdisciplinary research, these interoperability frameworks should be articulated in common ways and adopt global standards where relevant.'[50] As well as the international dimension, it is this vision of enabling community agreements around interoperability, and encouraging the identification and adoption of common standards, that drove the WorldFAIR project.

WorldFAIR worked with a set of eleven domain and cross-domain case studies[51], carefully chosen from existing CODATA and RDA activities or partnerships to provide maximum impact. Each case study (in project terms, a 'Work Package') developed an interoperability framework, comprising recommendations and/or a FAIR implementation for their discipline or interdisciplinary research area. Led by CODATA, a coordinating and synthesis activity (Work Package 2) supported each case study in understanding their

[34] Hodson, S., & Gregory, A. (2023). WorldFAIR Project (D1.3) 'First policy brief' https://doi.org/10.5281/zenodo.7853170
[35] https://worldfair-project.eu/
[36] Wilkinson, M., et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
[37] WorldFAIR included partners from the following countries: Belgium, Cyprus, Denmark, Germany, France, Ireland (EU); Norway (outside EU but eligible for Horizon framework funding); Australia, Brazil, Kenya, New Zealand, USA (outside EU and not usually eligible for EU funding); UK (at the start of the WorldFAIR project, the UK institutions were not eligible for EU funding).
[38] https://codata.org/
[39] https://council.science/
[40] https://www.rd-alliance.org/rda-europe
[41] https://www.nanocommons.eu/
[42] https://onegeochemistry.github.io/
[43] https://drexel.edu/lac/salurbal/overview/
[44] https://www.gbif.org/
[45] https://odis.iode.org/
[46] https://iupac.org/
[47] https://ddialliance.org/
[48] https://www.ohdsi.org/
[49] https://www.tdwg.org/
[50] European Commission, Directorate-General for Research and Innovation, Turning FAIR into Reality: final report and action plan from the European Commission expert group on FAIR data, Publications Office, 2018, https://data.europa.eu/doi/10.2777/1524; p. 29.
[51] https://worldfair-project.eu/case-studies-of-worldfair/

requirements through the completion of FAIR Implementation Profiles (FIPs). In turn these insights were incorporated into the development of a Cross-Domain Interoperability Framework (CDIF) and recommendations for future work on domain-sensitive FAIR guidelines and assessment.

Consequently, the perspective of the WorldFAIR project is avowedly global and holds that the EC, the European Research Area and EOSC have a major role to play as part of an international science ecosystem. It is firmly in the interest of the EU, and essential to the success of EOSC, to contribute positively to international data infrastructures, to international standards development, and the implementation of good practices as recommended in the CDIF.

POLICY ISSUES: EVIDENCE, ANALYSIS AND RECOMMENDATIONS

FAIR Data for 21st Century Science

The WorldFAIR policy recommendations reflect an international perspective, and a conviction regarding the need to orchestrate data internationally, to provide global scientific endeavours with the data needed to help address the planetary and societal challenges of the 21st Century. Science has an essential role in providing the evidence needed for effective policy-making and sustainable interventions. Open science and the FAIR principles are two related practices that can enable science to realise the opportunities afforded by the digital era. For this to be the case, transformations in scientific culture, methodologies, institutions, infrastructures and engagement with society are needed[52]. In this policy brief, we are primarily concerned with the necessary transformations that relate to the implementation of the FAIR principles as a means to equip scientific communities with the tools, infrastructure and data needed to pursue their mission in these changing circumstances, defined by the opportunities and affordances created by digital transformations.

1. Data Engineering

There is an urgent need for a shift from a 'bibliographic' data stewardship practice to a data engineering practice! The most fundamental recommendation to emerge from the WorldFAIR project is the following: to support the requirements of 21st Century science, we need to enable a transformation in our practice for data stewardship and move from a bibliographic approach to a data engineering approach.

In the bibliographic model, data is treated like a book in a library: a dataset is deposited in an appropriate domain specialist or generalist repository as a data package, with a persistent identifier and discovery metadata in an extended form of Dublin Core. This is, of course, better than nothing: the repository and the data stewards involved have performed an important service in ensuring that the data was not (to all intents and purposes) lost on a research group server or a personal hard drive. For such data to be reused, however, the dataset must be downloaded, and the significant task of data wrangling remains, often with inadequate, non-standard or only implicit information about the data and semantics. This is precisely the issue highlighted in the PWC report on the opportunity costs of not having FAIR data[53], and it falls well short of the EOSC and FAIR vision of machine-actionable data. If we persist with the bibliographic model, we will not achieve the 'web of FAIR data and services' promised by the EOSC.

In a seminal 2012 article 'Is Data Publication the Right Metaphor?'[54], Parsons and Fox drew our attention to this precise issue. Unfortunately, the 'data publication' or bibliographic model still dominates how we think about data stewardship, research data management, and the role of data repositories. A related issue is made clear by the WorldFAIR case studies and elsewhere: in many research fields too much effort and expense are

[52] These necessary transformations are well documented in the UNESCO Recommendation on Open Science https://unesdoc.unesco.org/ark:/48223/pf0000379949

[53] The PWC report for the European Commission estimated the opportunity cost of not having FAIR data to be 10.2 Bn per annum. The report also estimated that 'data cleansing' or data wrangling of poor quality data 'can take up to 80% of the total effort', leaving only 20% of project effort for analysis. European Commission, Directorate-General for Research and Innovation, Cost-benefit analysis for FAIR research data: cost of not having FAIR research data, Publications Office, 2019, https://data.europa.eu/doi/10.2777/02999

[54] Parsons, M.A. and Fox, P.A. (2013) 'Is Data Publication the Right Metaphor?', Data Science Journal, 12(0), p. WDS32-WDS46. Available at: https://doi.org/10.2481/dsj.WDS-042.

still going into extracting data, ex post facto, from academic articles or supplementary materials[55]. Data compilations are essential for effective research in many domains. At present, such efforts are hampered by numerous barriers. In many research areas (e.g. nanomaterials, plant-pollinator interactions) an astonishing amount of effort goes into the compilation of data extracted laboriously from published articles[56]. Consequently, essential data is often behind paywalls, and even more often communicated in formats that require a great deal of manual intervention. Such data should be deposited directly into the data aggregator as part of the workflow for making available ('publishing') the outputs from research and funded projects, and should be sufficiently well-described to support further use.

A significant part of the inefficiencies and opportunity costs identified in the PWC report, come from insufficient data stewardship, and they are left unaddressed by the data management involved in the bibliographic model. Quite simply, the use case in which researchers deposit data in support of their publication is not the only, or the most important, use case.

In the data engineering model, by contrast, data are treated as collections of 'datums' (individual data points, measurements, observations, annotations etc), each of which has sufficient associated information to be treated independently if necessary. This is what allows the selection of certain variables of interest, and their recombination with other variables in an aggregated resource or a new data product for analysis. In many fields, it is essential to integrate data from multiple sources to build up data on a topic globally to provide the evidence needed for well-grounded science and policy interventions.

Such data engineering is metadata-intensive and aims to create a network of (FAIR) data exchange. We understand the aspiration of EOSC to achieve a 'web of FAIR data and services' in precisely these terms. A good example of metadata facilitating a (relatively FAIR) data exchange is the use of SDMX[57] in the statistical and economic regulatory sector. Another example is GBIF, the Global Biodiversity Information Facility, which serves as the world's resource for biodiversity and species occurrence data[58]. Faced with an increasing variety of new data sources (including most notably environmental DNA and camera trap data), GBIF has been obliged to develop a new Unified Model that allows the integration of such data[59]. Another highly pertinent example is ODIS[60], the Ocean Data Information System, which uses an agreed architecture and a knowledge graph approach to federate data globally[61]. The data engineering paradigm is also evident in the work of INSPIRE to combine population health data with clinical outcome data, for research and policy uses[62].

The shift from a bibliographic approach to a data engineering approach is one of a magnitude which will necessitate considerable resourcing, investment, and upskilling, but which will also achieve significant benefits. The objective of such a shift is to provide a network for FAIR data exchange which will help to realise the opportunities noted above: automated data combination for cross-domain research and automated and fine-grained access control, ultimately better enabling the use of AI.

This cultural change will be a difficult one; nor is there likely to be a sufficient balance of cost-to-benefit for such an approach to be applied to all data, all the time. But we are witnessing increasing demand for research infrastructures to respond to the challenges described above. The use case that led to SDMX, GBIF and ODIS is pressing in many other research fields. Were this not the case, we would not see so many projects performing ex post facto extraction of data from research articles!

Furthermore, the energy, willingness, and requirement to undertake this transformation exists as a result of the FAIR vision, and it is one of the objectives of the Open Science movement and EOSC. Thus, from a policy perspective, we are faced with both significant challenges and a significant opportunity.

---

[55] In WorldFAIR, this is the case in a number of the research areas involved, but was particularly remarked on by the nanomaterials, geochemistry, biodiversity and plant-pollinator case studies.
[56] Drucker, D.P., et al. (2024). WorldFAIR (D10.2) 'Agricultural Biodiversity Standards, Best Practices and Guidelines Recommendations' https://doi.org/10.5281/zenodo.10666593, pp. 9-34, p. 26.
[57] Statistical Data and Metadata Exchange - https://sdmx.org/
[58] CODATA, the Committee on Data of the ISC, Pfeiffenberger, H., Uhlir, P., & Hodson, S. (2020). Twenty-Year Review of GBIF. Zenodo. https://doi.org/10.35035/ctzm-hz97;
[59] D9.2; https://worldfair-project.eu/biodiversity/
[60] https://odis.iode.org/
[61] https://worldfair-project.eu/ocean-science-sustainable-development/
[62] https://worldfair-project.eu/population-health/

Recommendation 1: Data Engineering.
Policy makers and funders need to encourage and enable a data engineering approach in the data infrastructures that are the most important to address major societal and planetary challenges. Specifically, this requires supporting long-lasting data aggregation and data integration services as part of EOSC and globally.

2. Metadata Uplift

An essential component of data engineering is the addition of **sufficiently detailed, standardised, and interoperable metadata**. To be crystal clear, here we are not referring simply to the metadata that enables discoverability, that helps address 'Findable' in the FAIR principles. We are referring to those components of metadata schema that facilitate interoperability, enabling data to be combined, integrated and reused. This includes, but is not limited to, definitions of variables (or quantities), descriptions of data structure, provenance and processing information[63]. DDI-CDI[64] and the draft CDIF profiles[65] provide examples of how this can be put into practice in cross-domain scenarios. In the DDI community, the term 'metadata uplift' has been coined to describe the task of adding sufficient information to facilitate analysis across datasets, and includes explorations of the use of machine learning (ML) to support that task[66].

For data to be (re-)used in research (i.e. compared, combined, analysed) the information about the data needs to be very detailed. The scientist — and their analytical tools — need to know what ***may*** be done with the data, and they need to know what ***can*** be done with the data. Cascading from this, knowing how the data can be processed, combined, and analysed requires — as a minimum — information about the concepts, variables and units involved, the data structure, the provenance and processing the data may have undergone, and estimates of their quality and accuracy. The increasing scale and complexity of research questions require an increasing scale of data stewardship and 'metadata uplift'.

While much attention is paid to artificial intelligence (AI) as the most important emerging technology, the factor which most limits its use, at least in science, is a mundane one: **research data that are not sufficiently described cannot be used by any form of intelligence, artificial or otherwise.** Much of the effort that goes into AI approaches, in relation to scientific discovery, is only necessary because the provenance and detail regarding the research data being consumed have been discarded after the initial data collection or generation. Good data stewardship, which preserves and publishes this detailed information about the data, will enable powerful AI techniques to realise their promise more fully, and support more effective use of traditional data analysis approaches as well.

Recommendation 2: Metadata Uplift.
Sufficiently detailed, standardised, and interoperable metadata are a precondition for the data products that are essential for high priority research areas. Research infrastructures and data infrastructures need to put into practice 'metadata uplift'. They must be enabled to do so by policy makers and funders.

[63] Here we employ the broad definition of 'metadata' that is accepted in many contexts and research disciplines. We recognise that some research fields think of these things (which include controlled vocabularies, ontologies, data models) as semantic artefacts, or as part of the data. We do not want to open this particular Pandora's box of terminological discord, but need to be clear that we are employing the term 'metadata' in its broadest sense.
[64] https://ddialliance.org/Specification/ddi-cdi
[65] Gregory, A., et al. (2024). WorldFAIR (D2.3) 'Cross-Domain Interoperability Framework (CDIF)' https://doi.org/10.5281/zenodo.11236871
[66] https://ddialliance.org/metadata-uplift-%E2%80%93-pdfexcel-to-structured-ddi-documentation; see also SRIA 1.2, p. 162 https://eosc.eu/wp-content/uploads/2023/12/20231114_SRIA_1.2_final2.pdf

There are four drivers for this call for data engineering and 'metadata uplift'. These are addressed in the following four recommendations and include:

1. Interdisciplinary research for global challenges;
2. Reproducible and transparent research;
3. Responsible use of AI in science;
4. Increasingly automated controlled access to sensitive data.

3. Interdisciplinary Research for Global Challenges

The major global human, societal, environmental and scientific challenges of our age are fundamentally interdisciplinary and related to all sectors of society. These challenges can only be addressed through the close collaboration of science, civil society, and government using cross-domain and multi-stakeholder research that seeks to understand complex systems through machine-assisted analysis, and enables data-driven decision-making processes[67]. The ability to combine data across traditional disciplines is becoming a sine qua non for many areas of research.

As argued above, this requires a shift to data engineering, enabled by the collaborative development, refinement and international adoption of sufficiently detailed, fine-grained, standardised, and interoperable metadata. Some examples of this have been provided above. Another is the EOSC Future 'Climate-Neutral and Smart Cities' Project[68] which combined European Social Survey data with climate and air quality data to create an important resource for understanding the relation of attitudes to climate change and environmental degradation based on lived experience. Notable about this work was the use of DDI-CDI, a CDIF component, to provide valuable information to researchers and data scientists about the way in which compound variables have been constructed. Data wrangling to create integrated data sets is frequently conducted as a one-off, labour intensive, manual process. The data engineering approach, and the call for 'sufficiently detailed, standardised, and interoperable metadata', seeks to establish the preconditions for such effort to be increasingly automated and repeated year on year for time series data. Doing so also with sufficient metadata about provenance and processing is foundational for reproducibility of findings and transparency.

Recommendation 3: Interdisciplinary Research for Global Challenges.
There is a need for investment in technologies and approaches that facilitate data aggregation and data integration for interdisciplinary, grand challenge research areas. Such investment should prioritise work that automates the integration approach and allows it to be performed year on year with time series data.

4. Reproducible and Transparent Research

Concerns about the reproducibility and replicability of scientific results have many dimensions. Although there are disputes about the scale of the 'crisis' and the way the phenomenon itself has been studied, perceptions that there ***is*** a crisis[69] are sufficiently strong for national and international initiatives to be launched to improve practice[70]. Universal application of the requirement that data be deposited in an appropriate repository — as a condition of publication — would help address a part of the issue[71]. Where such services exist in specific research fields, data should be deposited in a recognised regional or international data aggregator, or in a local or national service that can be harvested internationally[72].

Additionally, through the data engineering approach described above and explored in WorldFAIR, it is possible (in principle) to provide improved and more automated documentation of provenance and data lineage. Such information does not just pertain to how the data were created, but also to capturing — in machine-readable and machine-actionable ways — any processing (e.g. normalisation, data reduction),

[67] See for example, the International Science Council's Action Plan, area 2.1: https://council.science/actionplan/making-data-work-for-grand-challenges/
[68] https://eoscfuture.eu/data/climate-neutral-and-smart-cities/
[69] Baker, M. 1,500 scientists lift the lid on reproducibility. Nature 533, 452–454 (2016). https://doi.org/10.1038/533452a
[70] https://www.ukrn.org/global-networks/
[71] Miyakawa, T. No raw data, no science: another possible source of the reproducibility crisis. Molecular Brain, 13, 24 (2020). https://doi.org/10.1186/s13041-020-0552-2
[72] Drucker, D. P. et al. (2024). WorldFAIR (D10.3) 'Agricultural biodiversity FAIR data assessment rubrics' https://doi.org/10.5281/zenodo.10719265, p. 24-5.

integration and combination with other data, and analysis. To enable reuse, and to ensure transparency and reproducibility, it is important to be able to document (automatically) all the transformations data have undergone, as well as the methodologies, reasoning and logical steps used to arrive at conclusions.

Standards for describing data provenance — including PROV-O[73], the Common Workflow Language (CWL)[74], DDI-CDI Process descriptions[75], Validation and Transformation Language (VTL)[76], and Structured Data Transformation Language (SDTL)[77] — are all primarily descriptions of a stepwise process, involving actors, input/outputs, and processing actions. Shortcomings are encountered because the outputs of research are frequently used as inputs to further research, but the description of provenance (when it exists) is limited to a description of each stage in isolation. It is almost never the case that we can find a full description of data lineage using the existing standard models: they are implemented in a limited way, describing (typically) a single step in the chain. Existing models are insufficient for a complete description of data lineage, making it unrealistic to expect that data providers and users could provide this information. The need to connect such descriptions with each other to provide a fuller and machine-actionable summary of data lineage is emerging as an important area for attention.

Recommendation 4: Reproducible and Transparent Research.
The global data stewardship and metadata standards community should direct concerted effort to refine and improve standards for describing data provenance and processing, as well as technologies that enable such standards to be used to provide a full and machine-actionable account of data lineage. Such work should be enabled by funders and policy makers.

5. Increasingly Automated and Controlled Access to Sensitive Data

In many circumstances, access to data must be controlled. For many data archives, the manual management of access is a considerable time burden. Where this can be safely automated, efficiency benefits are achieved. A further related use case for the data engineering approach and increased automation is that of providing more fine-grained access to data sets (which may contain some sensitive data), and thereby addressing pressing issues of organisational and legal interoperability.

In systems in which entire files, or collections of files, are the object to which access rights are assigned, the 'access boundary' is necessarily limited to letting the most confidential and/or sensitive variables determine the access level to the whole package. For example, if we have a file containing 200 variables, of which two contain potentially sensitive, personal or disclosive information, then the entire file must be ruled off-limits for those who do not have sufficient access privileges. This phenomenon is termed 'data poisoning': that is, 198 of the 200 variables must be restricted (quarantined) for many legitimate users because of the 'toxic', disclosive nature of only two variables and through which they have been 'infected'. The real-world consequences of this are increased access times to data for end-users, and complex administrative overheads for data providers, with locally defined, often human-mediated workflows, which cannot scale up in the long term. **A significant limiting factor, therefore, is the inability to automate the negotiation of data access, which — while difficult — has been shown to be feasible in cases where sufficiently granular metadata exists.**

What we can understand from this is that the existing culture of data management, which often views the 'data set' as a monolithic entity acting as an adjunct to the publications it underpins, is a barrier to secondary data use. The data engineering approach, in which sets of datums, or selected variables, have sufficient associated information to be treated independently, provides a solution.

Some interesting steps along these lines are being taken in relation to the automation of access control, although these are less well developed. The UK Data Archive and others have been exploring how approaches borrowed from other domains can be combined with efforts in the world of research data to help make the granting of access to data more efficient, and also more 'focused' (that is, to be more permissive

[73] https://www.w3.org/TR/prov-o/
[74] https://www.commonwl.org/
[75] https://ddialliance.org/Specification/ddi-cdi
[76] https://sdmx.org/?page_id=5096
[77] https://ddialliance.org/products/sdtl/1.0

where 'data poisoning' has traditionally presented a barrier to access)[78]. These efforts show us that in order to automate access control, fine-grained management of data — at the variable level — as well as machine-actionable descriptions of the access conditions and roles and qualifications of the users given access, must be implemented.

It is a legal and moral imperative that access to data be granted in an ethical manner, in line with accepted best practice. At the same time, these obligations should not present any greater barrier to reuse of data than necessary. The key to finding the right balance — to increasing legal and organisational interoperability — is to have granular access control, so that data access can be as broad as possible, and no broader. Systems that can support negotiation of access to data in a dynamic and more automated way are needed to responsibly meet the demand for data while also generating efficiency benefits for data services.

Recommendation 5: Increasingly Automated and Controlled Access to Sensitive Data.
In partnership with the global data stewardship and metadata standards community, data services looking after sensitive data should direct concerted effort to developing and implementing systems that can support negotiation of access to data in a dynamic and more automated way. This effort should be supported by funders and policy makers.

6. Responsible Use of AI

Many of the same issues are relevant for the responsible use of AI in science. There is a lot of noise and churn on this topic, but the issues may be summarised as follows:

1. How may we best avoid the misuse of (sensitive) information in Large Language Models (LLMs) and other forms of AI?
2. How can we maximise transparency and reproducibility in results obtained using AI?
3. How may we best reduce the risk of hallucinations and imprecision?

How may we best avoid the misuse of (sensitive) information in LLMs and other forms of AI? The emergence of LLMs and generative AI have added extra complexity and potential risk to the issues described above in relation to sensitive data. The concerns are twofold: first, that the new technologies increase the risk, for example, of accidental disclosure and identification; and second, that the increased risk will make organisations more risk-averse and throttle progress on machine accessibility to data.

It is always important to describe what the data can be used for, and who is allowed to use it. When machines become an important class of data users — especially at the scale of LLMs — it is essential that the terms of use, and the qualifications required for use, be made available to them in a machine-actionable fashion. For research infrastructures, particularly those dealing with sensitive data, it will be important to explore the extent to which contracts, licences, the criteria for access, the qualifications of the potential user, and the legislation that underpins each of these things, can be made machine-actionable. There is currently no good way to describe terms and conditions of use for data in a standard, machine-actionable way, although frameworks such as the Open Digital Rights Language (ODRL)[79] are a promising start, as is work such as the Data Privacy Vocabulary (DPV)[80], both from W3C. Neither is sufficient to meet the current challenge, however, in their current form. Further work is necessary on this topic[81].

How can we maximise transparency and reproducibility in results obtained using AI? The shortcomings of current models to describe data lineage and the need for further work on this topic has been described above. AI — and LLM-based generative AI in particular — makes this situation even worse. These technologies present us with 'black boxes' which do not act in a stepwise fashion suitable for description using typical approaches. There is no visibility into the process other than through an

[78] Lungley, Deidre, Gilders, Thomas, Bell, Darren, & Rumiancev, Artiom. (2022, November 30). Towards Machine-assisted Disclosure Assessment with DDI-CDI, DPV and sdcMicro. EDDI2022: The 14th Annual European DDI User Conference (EDDI2022), Sciences Po, Paris, France. https://doi.org/10.5281/zenodo.7656053
[79] https://www.w3.org/TR/odrl-model/
[80] https://www.w3.org/community/dpvcg/
[81] Rouchon, O., et al. (2024). D6.2 - Core metadata schema for legal interoperability' https://doi.org/10.5281/zenodo.11104269 for a discussion of some of these issues and a survey of metadata profiles.

examination of the algorithms and the code which implements it, informed by a knowledge of the training set used. Such information is rarely available[82].

The DDI-CDI process model may help in this regard and indicate a useful direction for further work. The model supports the case where the inputs, success criteria, and a 'playbook' of functions can be described, along with the processing engine. This information then becomes part of the provenance description attached to the resulting data. While still not providing visibility into the operations of the black box, it does give a reasonable set of information for use in determining who did what to any given set of data.

How may we best reduce the risk of hallucinations and imprecision? 'Hallucinations' are a sufficiently known phenomenon with generative AI, for the term to have become a commonplace. More dangerous are those cases where an unknown degree of imprecision applies to results which seem plausible on the surface. The key point of discussion here tends to be around the data used to train LLMs and other AI models. Within the FAIR community, discussions are underway to explore how a better description of data, along the lines of the FAIR Principles, can help to protect society from the worst tendencies of AI, by diminishing the likelihood of hallucinations and imprecision[83]. Google and others have been developing the Croissant[84] specification for providing metadata descriptions of training data sets: these models are similar to (but semantically poorer than) the DDI-CDI standard advocated in CDIF, but they are currently compatible (generating a Croissant description from a DDI-CDI one would be relatively trivial). The idea that accurate, structured metadata could help reduce imprecision and hallucination does, however, appear to be a sound assumption and worthy of further exploration.

Recommendation 6: Responsible Use of AI.
AI technologies (particularly generative tools based on LLMs) present a number of challenges, notably the potential misuse of sensitive information, a lack of transparency and reproducibility, and the risk of hallucinations and imprecision. The use of detailed, accurate and structured metadata should be explored as one of the means of enhancing the utility and precision of these technologies and to help in imposing guardrails.

The remaining recommendations outline actions to enable data engineering and metadata uplift, and provide the tools to respond to the challenges in the four drivers just discussed.

7. Support the Further Development of the Cross-Domain Interoperability Framework (CDIF)

The CDIF idea emerges from several years of discussion and related work and is conceived as a set of recommendations for the coordinated use of standards to enable automated, cross-domain sharing of data and metadata. By establishing a 'lingua franca' based on existing cross-domain standards, it will become possible to implement cross-domain exchange of FAIR metadata and data — including data discovery, access, and integration — in a scalable fashion. In the simplest terms, CDIF is a compilation of recommended practices for implementing relevant standards, to support the needed interchange of data and metadata. The first set of CDIF modules[85] make recommendations for the use of cross-domain standards and practices for 1) discovery (of data and metadata resources); 2) data access (specifically, machine-actionable descriptions of access conditions and permitted use); 3) controlled vocabularies (good practices for the publication of controlled vocabularies and semantic artefacts); 4) data integration (description of the structural and semantic aspects of data to make it integration-ready); and 5) universals (the description of 'universal' elements, time, geography, and units of measurement). Five areas of future work are also discussed: i) provenance (the description of provenance and processing); ii) context (the description of 'context' in the form of dependencies between fields within the data and a description of the research setting); iii) AI (discussing the impacts of AI and the role that metadata can play); iv) packaging (the creation

[82] Research is moving rapidly. Approaches are being explored through which it can be indicated when specific nodes within the knowledge graph driving the LLM have been used in the response to a prompt: https://arxiv.org/html/2312.16374v2.
[83] https://www.lorentzcenter.nl/the-road-to-fair-and-equitable-science.html
[84] Akhtar, M, et al., (2024), 'Croissant: A Metadata Format for ML-Ready Datasets' https://doi.org/10.48550/arXiv.2403.19546
[85] Gregory, A., Bell, D., Brickley, D., Buttigieg, P. L., Cox, S., Edwards, M., Doug, F., Gonzalez Morales, L. G., Heus, P., Hodson, S., Kanjala, C., Le Franc, Y., Maxwell, L., Molloy, L., Richard, S., Rizzolo, F., Winstanley, P., & Wyborn, L. (2024). WorldFAIR (D2.3) (Version 1). Zenodo. https://doi.org/10.5281/zenodo.11236871

of archival and dissemination packages); v) additional Data Formats (support for some of the data formats not fully supported in the initial release, such as NetCDF[86], Parquet[87], and HDF5[88]).

A number of projects to implement and test the CDIF recommendations are underway, scheduled or planned (notably under the banner of WorldFAIR+). We invite collaboration and feedback to refine, improve and extend the CDIF profiles and recommendations.

Recommendation 7: Support the Further Development of the Cross-Domain Interoperability Framework (CDIF).
Given the pressing need for a date engineering approach to support interdisciplinary research, policy makers and funders should enable the further development of CDIF and its extension to more profiles and recommendations. The data stewardship community, metadata standards organisations and data infrastructures should collaborate to implement, test, refine and improve the current and forthcoming CDIF profiles and recommendations.

8. Invest in Research Infrastructures

The proposed shift from a 'bibliographic' data stewardship practice to a data engineering practice is one of magnitude which will require considerable resourcing, investment, and upskilling; but it is one which will also achieve significant benefits. Increasingly, one of the principal roles of research infrastructures will be to provide researchers with pre-integrated, 'science-ready', interdisciplinary data products.

Few research infrastructures are currently equipped to face this task. Concern will understandably be expressed by many data archives and repositories that are under-resourced even for the current tasks of providing file-level metadata and digital preservation. By the same token, there is clearly considerable demand for this transformation and ambition on the part of many research infrastructures to be able to combine datasets for interdisciplinary research.

An excellent example of this, already mentioned above, is the EOSC Future 'Climate-Neutral and Smart Cities' project[89]. While a part of this project looked at the detailed aspects of data integration with data coming from different sources, it also examined the organisational and scientific aspects of collaboration. Research infrastructures collaborate in what is too often an ad hoc fashion, with no planning around what other data resources could most usefully be combined with their own from a scientific perspective. This lack of preparation is seen in how resources are allocated, and how scientific projects are supported. It may not be reasonable to expect a domain-focused data repository to address cross-domain requirements, but there is a need for a collaborative framework which will make organisational and scientific collaboration a more normal part of their functioning[90].

Such cross-domain and cross-infrastructure collaboration is not possible unless the data to be integrated is well-described in the manner suggested above: the data engineering approach is a precondition for making collaborative interdisciplinary science practical. In fact, much of the technology needed for more granular data management already exists. Many data archives have implemented 'shopping basket' applications where researchers can select a set of variables from across the various data sets held by the archive, and such implementations reflect a change in the way these technologies are implemented rather than a shift in the underlying technology.

This change is characterised by data management at the variable level, instead of solely focusing on the data set as a monolithic entity. Within WorldFAIR, we can see this in practice in the SALURBAL case[91], in the

[86] https://en.wikipedia.org/wiki/NetCDF
[87] https://parquet.apache.org/
[88] https://www.hdfgroup.org/solutions/hdf5/
[89] https://eoscfuture.eu/data/climate-neutral-and-smart-cities/
[90] https://preprints.arphahub.com/article/115047/
[91] Quistberg DA, et al. 'Building a Data Platform for Cross-Country Urban Health Studies: the SALURBAL Study'. J Urban Health. 2019 Apr;96(2):311-337. doi: https://doi.org/10.1007/s11524-018-00326-0. PMID: 30465261; PMCID: PMC6458229.

INSPIRE Network case[92], and elsewhere. Many data infrastructures are beginning to provide this type of capability (e.g. ODIS[93] and GBIF[94] which are also partners in the WorldFAIR project).

Given the preceding observations, it is essential to insist on two points. First, the transition from bibliographic data stewardship to data engineering categorically should not place an extra burden on researchers. On the contrary, this transition, if realised, will reduce that burden. Second, those infrastructures and data services that are involved in providing researchers with data (e.g. large facilities, planetary observing systems, social surveys and so on) should be adequately resourced to provide FAIR data, to serve fully integrated data products that correspond to researchers' needs. It is through sufficient and ambitious resourcing and implementation of a data engineering approach described here that we can address the opportunity cost of not having FAIR data and reduce the amount of effort spent on data wrangling. The same applies to the so-called 'long tail' of research data production, including data creation or collection activities in research performing organisations, from large faculties to individual researchers: the objective should not be to turn researchers into data stewards, but to make it easier for data stewards to furnish researchers with the data they need for analysis and knowledge creation.

Recommendation 8: Invest in Research Infrastructures.
Research infrastructures should be encouraged and enabled to transition to a data engineering approach and provide researchers with pre-integrated, 'science-ready', interdisciplinary, data products. This will require investment to build capacity and to adopt new technologies and approaches. This transition should be enabled by funders and policy makers. It should be embraced by research infrastructures and data repositories as a necessary evolution of their function and mission.

9. Enable FAIR Practice in Research Communities

Technical standards that can communicate information about all aspects of data need to exist and be implemented, both at the level of domains ***and*** for the purposes of cross-domain exchange. Many research disciplines have developed technical standards for metadata, and these are to varying extents adopted. While some domains are well organised regarding metadata exchange, others rely on the organic emergence of good practices which remain relatively ungoverned or are dominated by the implementations of major infrastructure players. It is essential to enable research disciplines to develop data standards, to support their implementation, and to encourage good practice.

A singular feature of the WorldFAIR strategy and approach was to emphasise the importance of the role played by various stakeholders that can genuinely represent a research discipline, or a significant part of it. It is such organisations that can pose the question 'how do we want our science to be conducted?' and thereby articulate the wishes of a community. The WorldFAIR project was designed to involve a number of such organisations. Particularly important in this role are longstanding International Scientific Unions like IUPAC (founded in 1919), or emergent initiatives such as OneGeochemistry.

Similarly, those standards organisations that serve particular research communities (the Observational Health Data Sciences and Informatics [OHDSI] program in clinical medical research, the DDI Alliance in social, behavioural and economic sciences, etc.) as well as more general standards bodies serving broader needs (such as the W3C for Web standards) must be engaged.

The WorldFAIR project demonstrated the importance of a multi-stakeholder approach and the role that can be played by a number of different types of organisations and stakeholders in advancing FAIR practice. Such organisations include: global infrastructures (GBIF, WP09; ODIS, WP11); international representative and authoritative organisations (IUPAC, WP03; OneGeochemistry, WP05); data infrastructures (Sikt and ADA, WP06; INSPIRE, WP07); metadata standards organisations (IUPAC, WP03; DDI, WPs 06, 07, 08; TDWG, WP 09 and 10; GloBI, WP10); and influential projects (NanoCommons, WP04; INSPIRE, WP07; SALURBAL, WP08).

[92] INSPIRE PEACH is based on the OMOP Common Data Model, which provides for granular specification of data for reuse: https://aphrc.org/inspire/project/a-platform-for-evaluation-and-analysis-of-covid-19-harmonised-data-peach-2/
[93] https://oceaninfohub.org/odis/
[94] https://www.gbif.org/

Recommendation 9.1: Enable standards bodies and international and representative organisations.
To advance FAIR implementation, it is important to take a multi-stakeholder approach. In particular, policy makers and funders should help support the role played by representative and authoritative organisations as well as standards bodies on an international scale.

Recommendation 9.2: Support WorldFAIR+.
The WorldFAIR approach should be further tested and refined as a means of enabling research communities to develop and implement good practice, through a multi-stakeholder approach. Policy makers and funders should support the WorldFAIR+ initiative and similar activities.

As recommended in the 'Turning FAIR into Reality' report, "Research communities need to be supported to develop interoperability frameworks that define their practices for data sharing, data formats, metadata standards, tools and infrastructure."[95] Similarly, the UNESCO Recommendation on Open Science calls on Member States and other stakeholders to promote "Community agreements, concluded in the context of regional or global research communities, and which define community practices for data sharing, data formats, metadata standards, ontologies and terminologies, tools and infrastructure. International scientific unions and associations, regional or national research infrastructures and journal editorial boards each have a role to play in helping develop these agreements. In addition, convergence between the various semantic artefacts (particularly vocabularies, taxonomies, ontologies and metadata schema) is essential for the interoperability and reuse of data for interdisciplinary research."[96] The FIPs exercises undertaken by the WorldFAIR project are a useful approach to enable communities to articulate, reflect on, and evolve their practice. The FIPs approach will be more effective if it is conducted internationally at a global scale. Both CODATA (through WorldFAIR+ and its collaborations with the UN System and International Scientific Unions) and the Research Data Alliance (through its Communities of Practice) are well-placed to coordinate such activities.

Recommendation 9.3: Support the FIPs approach and infrastructure.
Funders and policy makers, through investment and appropriate policy direction, should support the global data stewardship community to socialise the FIPs approach, and to improve and sustain the necessary infrastructures.

Recommendation 9.4: Support CODATA and RDA to enable research communities.
The Research Data Alliance and CODATA should be supported in their activities to enable research disciplines and interdisciplinary research areas to advance their FAIR practices.

10. Support the Sustainability of Semantic Artefacts

Many of the organisations that develop and maintain specifications and standards are international and global in scope. These include treaty organisations like BIPM[97], large and well-resourced standards organisations like the Open Geospatial Consortium[98], as well as smaller entities serving particular domains, like the DDI Alliance[99] or TDWG[100]. Some International Scientific Unions like IUPAC[101] have a particular vocation to establish and maintain standards and terminologies. Some of these entities are well-financed. Others rely extensively on volunteer effort, or time donated by researchers and data experts. Although such activities can be extremely effective, they are often significantly under-resourced, particularly when the passion of creating a resource gives way to the more mundane (and less visible) task of maintaining it. The challenges involved in maintaining or transitioning technical platforms, handing over leadership, or simply of financial sustainability are present for many such initiatives. Such entities have a wide range of business models, governance arrangements and juridical statuses.

[95] European Commission, Directorate-General for Research and Innovation, Turning FAIR into reality: final report and action plan from the European Commission expert group on FAIR data, Publications Office, 2018, https://data.europa.eu/doi/10.2777/1524; p.29.
[96] UNESCO Recommendation on Open Science, VI.iii.18.f, p. 24. https://unesdoc.unesco.org/ark:/48223/pf0000379949
[97] https://www.bipm.org/en/
[98] https://www.ogc.org/
[99] https://ddialliance.org/
[100] https://www.tdwg.org/
[101] https://iupac.org/

Standards such as Schema.org[102], DCAT[103], SKOS[104], PROV-O[105], and DDI-CDI[106] are recommended in the draft CDIF. These have a number of different origins and maintainers, but the prominence of Web standards developed and maintained under the auspices of W3C is notable. The stability and sustainability of standards included in the CDIF will be important, and a matter of interest for EOSC and other stakeholders internationally.

The availability of funding to ensure the further development of such entities, even when they are outside the EC (on the model that led to WorldFAIR) would be extremely beneficial to European research and to EOSC. If international funding partnerships can be agreed, similar in ambition to the Belmont Forum, to assist data and metadata standards development, so much the better; but the EC has the capacity and mission to lead the way. One immediate step could be a survey and analysis of business models, governance and juridical statuses along the lines of the study CODATA previously conducted with OECD on 'Business Models for Sustainable Data Repositories'[107].

Recommendation 10: Support the Sustainability of Semantic Artefacts.
Policy makers and funders should help to ensure the sustainability of the standards organisations that develop and maintain metadata schema, terminologies, and other semantic artefacts. An immediate step would be to fund a project to conduct survey and analysis of business models, governance and juridical status, and make recommendations to improve the sustainability and effectiveness of such organisations.

11. Strengthen International Partnerships

True to its remit and mission, the WorldFAIR project takes a global view and insists that the EC, the European Research Area, and EOSC have a major role to play as part of an international science ecosystem. It is not a mere truism to observe that science is a global endeavour, with global ramifications. Similarly, standardisation efforts — particularly in science, data and metadata — need to be global. It is firmly in the interest of the EU, and essential to the success of EOSC, to contribute positively to international data infrastructures and standards development.

Therefore, the creation of partnerships with authoritative international organisations or initiatives is a necessary part of building EOSC and should not be neglected. To avoid building a silo, EOSC should avoid the temptation to 'go it alone'. On the contrary, EOSC should actively partner with existing international data infrastructures and organisations that maintain standards and metadata schema (particularly those related to the UN data ecosystem), with inter-governmental initiatives and with global representative bodies such as the International Science Council and the International Scientific Unions. The Research Data Alliance and CODATA have important roles to play to assist with the internationalisation of EOSC activities.

Recommendation 11: Strengthen International Partnerships.
EOSC should actively build international partnerships with standards bodies, inter-governmental initiatives and global representative bodies. Contributing to the development of a strong global ecosystem for data and metadata exchange will benefit EOSC and is necessary for its success.

[102] https://schema.org/
[103] https://www.w3.org/TR/vocab-dcat-3/
[104] https://www.w3.org/2004/02/skos/
[105] https://www.w3.org/TR/prov-o/
[106] https://ddialliance.org/Specification/ddi-cdi
[107] OECD (2017), "Business models for sustainable research data repositories", OECD Science, Technology and Industry Policy Papers, No. 47, OECD Publishing, Paris, https://doi.org/10.1787/302b12bb-en

POLICY RECOMMENDATIONS

Recommendation 1: Data Engineering
Policy makers and funders need to encourage and enable a data engineering approach in the data infrastructures that are the most important to address major societal and planetary challenges. Specifically, this requires supporting long lasting data aggregation and data integration services as part of EOSC and globally.

Recommendation 2: Metadata uplift
Sufficiently detailed, standardised, and interoperable metadata are a precondition for the data products that are essential for high priority research areas. Research infrastructures and data infrastructures need to put into practice 'metadata uplift'. They must be enabled to do so by policy makers and funders.

Recommendation 3: Interdisciplinary research for global challenges
There is a need for investment in technologies and approaches that facilitate data aggregation and data integration for interdisciplinary, grand challenge research areas. Such investment should prioritise work that automates the integration approach and allows it to be performed year on year with time series data.

Recommendation 4: Reproducible and transparent research
The global data stewardship and metadata standards community should direct concerted effort to refine and improve standards for describing data provenance and processing, as well as technologies that enable such standards to be used to provide a full and machine-actionable account of data lineage. Such work should be enabled by funders and policy makers.

Recommendation 5: Increasingly automated and controlled access to sensitive data
In partnership with the global data stewardship and metadata standards community, data services looking after sensitive data should direct concerted effort to developing and implementing systems that can support negotiation of access to data in a dynamic and more automated way. This effort should be supported by funders and policy makers.

Recommendation 6: Responsible use of AI
AI technologies (particularly generative tools based on LLMs) present a number of challenges, notably the potential misuse of sensitive information, a lack of transparency and reproducibility, and the risk of hallucinations and imprecision. The use of detailed, accurate and structured metadata should be explored as one of the means of enhancing the utility and precision of these technologies and to help in imposing guardrails.

Recommendation 7: Support the further development of the Cross-Domain Interoperability Framework (CDIF)
Given the pressing need for a date engineering approach to support interdisciplinary research, policy makers and funders should enable the further development of CDIF and its extension to more profiles and recommendations. The data stewardship community, metadata standards organisations and data infrastructures should collaborate to implement, test, refine and improve the current and forthcoming CDIF profiles and recommendations.

Recommendation 8: Invest in Research Infrastructures
Research infrastructures should be encouraged and enabled to transition to a data engineering approach and provide researchers with pre-integrated, 'science-ready', interdisciplinary data products. This will require investment to build capacity and to adopt new technologies and approaches. This transition should be enabled by funders and policy makers. It should be embraced by research infrastructures and data repositories as a necessary evolution of their function and mission.

Recommendation 9.1: Enable standards bodies and international and representative organisations
To advance FAIR implementation, it is important to take a multi-stakeholder approach. In particular, policy makers and funders should help support the role played by representative and authoritative organisations as well as standards bodies on an international scale.

Recommendation 9.2: Support WorldFAIR+
The WorldFAIR approach should be further tested and refined as a means of enabling research communities to develop and implement good practice, through a multi-stakeholder approach. Policy makers and funders should support the WorldFAIR+ initiative and similar activities.

Recommendation 9.3: Support the FIPs approach and infrastructure
Funders and policy makers, through investment and appropriate policy direction, should support the global data stewardship community to socialise the FIPs approach, and to improve and sustain the necessary infrastructures.

Recommendation 9.4: Support CODATA and RDA to enable research communities
The Research Data Alliance and CODATA should be supported in their activities to enable research disciplines and interdisciplinary research areas to advance their FAIR practices.

Recommendation 10: Support the sustainability of semantic artefacts
Policy makers and funders should help to ensure the sustainability of the standards organisations that develop and maintain metadata schema, terminologies, and other semantic artefacts. An immediate step would be to fund a project to conduct survey and analysis of business models, governance and juridical status, and make recommendations to improve the sustainability and effectiveness of such organisations.

Recommendation 11: Strengthen international partnerships
EOSC should actively build international partnerships with standards bodies, inter-governmental initiatives and global representative bodies. Contributing to the development of a strong global ecosystem for data and metadata exchange will benefit EOSC and is necessary for its success.

RECOMMENDATIONS FROM WORLDFAIR SYNTHESIS REPORTS

The WorldFAIR Synthesis and Coordination work package, WP02, produced — at the end of the project — three reports that include recommendations in relation to the use of FAIR Implementation Profiles (FIPs), the Cross-Domain Interoperability Framework and FAIR assessment. These recommendations are presented below.

Recommendations in relation to FIPs

What follows are the most important recommendations that have emerged from the WorldFAIR experience of using FAIR Implementation Profiles (FIPs) as presented in the report 'WorldFAIR's Experience with FIPs'[108]. The recommendation type and its intended audience are specified in brackets after each recommendation.

1. **Provide policy support and investment for FIPs.** Funders and policy makers, through investment and appropriate policy direction, should support the global data stewardship community to socialise the FIPs approach, and to improve and sustain the necessary infrastructures. (Policy; policy makers and funders).

2. **Publish FERs:** Representative and authoritative organisations that are responsible for particular FAIR Enabling Resources (FERs) should describe and publish these FERs in a way that they can be easily referenced in FIPs. They should be supported and enabled in doing so by other stakeholders (funders, policy makers and the data stewardship community at large). (Organisational; authoritative/representative organisations, standards organisations, data repositories).

3. **Make it easier to publish FERs and FIPs:** Supported by the global data stewardship community, the GO FAIR Foundation (GFF) should explore solutions to lower the barriers to information entry to the FIPs ecosystem, including by ensuring a wider range of import and export profiles. The GFF should be enabled to do so by well-targeted funding to enable the development and sustainability of the necessary infrastructures. (Organisational; GFF, data stewardship community, funders, policy makers).

    a. Specifically, JSON-LD import and export should be developed and enabled in the FIPs Wizard and FAIR Connect.[109]

    b. Specifically, the GFF (and any other stakeholders using systems that will interoperate) should enable a distinction to be made between 'no answer given' (e.g. due to lack of maturity of the community) and 'not applicable' (e.g. because the community has determined that no declaration is necessary, no resource is needed for this sub-principle).

4. **Make it easier to visualise FIPs:** Collectively, the global data stewardship community and the GFF should explore numerous avenues to improve the way in which FIPs and FERs can be visualised and analysed. This endeavour should be enabled by well-targeted funding and direction from policy makers. (Organisational; GFF, data stewardship community, funders, policy makers).

5. **Enable the wholesale use of FAIRsharing entries as FERs in FIPs:** GFF and FAIRsharing, supported by the global data stewardship community and other stakeholders, should explore mechanisms to enable the wholesale use of FAIRsharing information as FERs in FIPs. Such a process should be supported by well-targeted funding and direction from policy makers. (Technical; GFF, FAIRsharing, data stewardship community, funders, policy makers).

[108] Hodson, S. (2024). WorldFAIR (D2.2) WorldFAIR's Experience with FIPs (second set of FAIR Implementation Profiles for each case study) (Version 1). Zenodo. https://doi.org/10.5281/zenodo.11236094

[109] Since the draft version of this report was shared with colleagues at the GFF, a JSON-LD export has been enabled for the FIP Wizard.

6. **Review optimal encoding of FIPs and explore alignment:** GFF, ODIS, the CDIF WG and AG, and global data stewardship community, should constructively review the pros and cons of the technical underpinnings to the current FIPs approach and explore whether an alignment of the nanopublications approach and the suggested JSON-LD/schema.org approach could be achieved. Such a process should be supported by well-targeted funding and direction from policy makers. (Technical; GFF, FAIRsharing, data stewardship community, funders, policy makers).

Recommendations in relation to CDIF

The WorldFAIR report on the Cross-Domain Interoperability Framework (CDIF)[110] **presents a core set of five CDIF profiles**, which address the most important functions for cross-domain FAIR implementation.

1. **Discovery** (discovery of data and metadata resources);
2. **Data access** (specifically, machine-actionable descriptions of access conditions and permitted use);
3. **Controlled vocabularies** (good practices for the publication of controlled vocabularies and semantic artefacts);
4. **Data integration** (description of the structural and semantic aspects of data to make it integration-ready);
5. **Universals** (the description of 'universal' elements, time, geography, and units of measurement).

Each of these profiles is supported by specific **technical recommendations**, including the set of metadata fields in specific standards to use, and the method of implementation to be employed for machine-level interoperability.

A further set of topics is examined, establishing the priorities for further work. These include:

1. **Provenance** (the description of provenance and processing);
2. **Context** (the description of 'context' in the form of dependencies between fields within the data and a description of the research setting);
3. **Perspectives on AI** (discussing the impacts of AI and the role that metadata can play);
4. **Packaging** (the creation of archival and dissemination packages);
5. **Additional Data Formats** (support for some of the data formats not fully supported in the initial release, such as NetCDF[111], Parquet[112], and HDF5[113]).

In each of these topics, current discussions are documented, and considerations for further work are provided. The **technical recommendations** related to each profile may be summarised as follows:

General: CDIF metadata should be embedded in landing pages or linked stand-alone files, encoded in JSON-LD. The supported profiles will be indicated as part of the metadata.

Discovery profile: This profile recommends the use of a set of key Schema.org fields for describing static data sets and queryable data sources, with the DCAT[114] equivalent recognised as an acceptable alternative.

Access profile: This profile recommends that ODRL[115] Actions and Entities be used to describe policies and conditions for the use of data. At this time, the utility of this approach is limited by the lack of shared vocabularies for conditions of use, user qualifications, legal constraints, and similar important items. ODRL is thus limited to describing policies in terms of the disseminating institution but provides a basis for expansion in future when the needed vocabularies are developed.

Controlled vocabularies profile: This profile recommends the use of SKOS[116] for describing controlled vocabularies, understood to mean any terminological resource. The use of OWL[117] as a linked extension to

[110] Gregory, A., et al. (2024). WorldFAIR (D2.3) https://doi.org/10.5281/zenodo.11236871
[111] https://en.wikipedia.org/wiki/NetCDF
[112] https://parquet.apache.org/
[113] https://www.hdfgroup.org/solutions/hdf5/
[114] https://www.w3.org/TR/vocab-dcat-3/
[115] https://www.w3.org/TR/odrl-model/
[116] https://www.w3.org/TR/2009/REC-skos-reference-20090818/
[117] https://www.w3.org/TR/owl2-overview/

what is presented in SKOS is also recommended, as is the use of XKOS[118] for formal statistical classifications.

Data description profile: This profile recommends the use of DDI-CDI[119] to provide a granular description of the structure of data sets, and how the logical content of those datasets relates to their physical encoding. Text-based data is supported (CSV and other delimited formats, fixed-width ASCII, etc.), with the intention of expanding support for other types of data in future. The recommendations cover description of individual data sets to make them 'integration-ready'.

Universals profile: This section recommends the information which should be provided when describing time, geography, and units of measurement in other metadata sets. Some standards for this purpose are recommended in each area.

Recommendations in relation to FAIR assessment

Presented in the WorldFAIR report D2.4 'Recommendations and framework for FAIR Assessment within (and across) disciplines'[120], the key recommendation on FAIR assessment from WorldFAIR is to 'put the FIP horse before the FAIR assessment cart'. Specifically, this means that it is essential to encourage and enable research communities to reflect on their practice through developing FIPs. Those FIPs can then be used as the basis for FAIR assessment in that research domain.

Our survey of FAIR assessment tools indicates that most issues arise when a 'one-size-fits-all' tool does not deal sufficiently well with domain practices. Therefore, FAIR assessment tools must consider the community practice and convergence on recognised FAIR standards and technologies within those communities. One of the best guides we have to do this is the FIPs methodology.

1) **Funders and policy makers should encourage and enable research communities to develop FAIR Implementation Profiles so that these can be used as the basis of alignment and ultimately assessment.** (Policy recommendation; for funders and policy makers, research communities).

In the development and use of FAIR assessment tools, it is essential to be clear about their purpose. This includes being precise about what is being assessed (e.g. the FAIR-supporting practices of repositories, or the FAIRness of data sets) and in relation to what entity the assessment is being made (e.g. repository, dataset, community).

Realising practical interoperability requires that we have agreements within communities regarding FAIR practice. Assessment of FAIRness requires that we are able to formalise such agreements in a useful way and incorporate them into our metrics.

The role of standards-setting organisations of various sorts is essential in setting benchmarks against which FAIR assessment can be conducted. This includes publishing FERs and FIPs.

2) **Funders and policy makers should encourage and enable authoritative organisations — those that can meaningfully represent specific research communities — to formalise agreements on FAIR practice, in particular by ensuring FERs are published and that the community curates FIPs to describe its practice.** (Policy recommendation; for funders and policy makers, research communities).

Finally, after discussing some of the challenges of FAIR assessment in relation to domain-specific and cross-domain requirements, including those of machine-to-machine interoperability, we present the following recommendations for a framework for FAIR assessment:

[118] https://ddialliance.org/Specification/RDF/XKOS
[119] https://ddialliance.org/Specification/ddi-cdi
[120] Gregory, A., & Hodson, S. (2024). WorldFAIR (D2.4) 'Recommendations and framework for FAIR Assessment within (and across) disciplines'. https://doi.org/10.5281/zenodo.11242737

A. **A framework for FAIR assessment should include domain-agnostic and domain-specific criteria. Cross-domain criteria form a third category in certain use cases.** (Technical recommendation; for developers of FAIR assessment criteria and tools).

B. **A framework for FAIR assessment should take into account the practices expressed in specific disciplines: FAIR assessment should be weighted accordingly, and the information should be used to improve guidance.** (Technical recommendation; for developers of FAIR assessment criteria and tools).

C. **To include machine-to-machine interoperability, a framework for FAIR assessment will need to reference profiles for how FERs are implemented. CDIF shows us the level of specificity which may be necessary to achieve machine-to-machine interoperability. Developments in the standards space in relation to 'shapes' and 'profiles' provide an indication for how this can be done and direction for future work.** (Technical recommendation; for developers of FAIR assessment criteria and tools).

SUSTAINABILITY AND LEGACY

WorldFAIR's plans to sustain outputs and to accentuate the outcomes of the project are detailed comprehensively in D14.2 'WorldFAIR Sustainability and Exploitation Plans'[121].

WorldFAIR's strategy, in design and conduct, was for each of the case studies to involve either i) representative, authoritative organisations, at international scale, or ii) significant projects or collaborations. This being the case, all the outputs will continue to be used and refined and the outputs embedded in the activities of the partners and their communities. All of the Case Study outputs are of utility and interest to research communities involved in EOSC. Some of the WPs will continue as RDA Communities of Practice or as CODATA groups or both.

As described above, it is intended that the WorldFAIR synthesis work make a direct contribution to EOSC. Most notably, the CDIF is a significant milestone that provides guidance and recommendation for implementers to address interoperability and reusability issues for cross-domain research. The recommendations for FIPs and FAIR assessment are of direct relevance for EOSC and related projects. Above all we hope that the policy recommendations will be heeded and EOSC will promote a data engineering approach in order to build the web of FAIR data and services.

To ensure this contribution, the CDIF work will continue, coordinated by CODATA. The Working Group and Advisory Group will be maintained and expanded. Next steps include 1) breaking down the current profiles into stepwise actions for implementers, supported by code and examples; and 2) working on the next set of CDIF profiles, including provenance / data lineage and further support for data integration.

The objectives of WorldFAIR, the methodology, the case study approach and the focus on cross-domain challenges was conceived before the project and will continue after it. As part of the 'Making Data Work for Cross-Domain Grand Challenges' programme of activity to help deliver a specific part of the International Science Council (ISC) Action Plan with the same name, CODATA will expand and sustain the vision and methodology being advanced through the WorldFAIR Project.

WorldFAIR+ is conceived as a federation of aligned projects, providing case studies that will further test and refine the WorldFAIR methodology. New case studies in emergencies and geosciences will start in the second half of 2024. The CDIF Recommendations and the work of Work Package 7, will be implemented in a new Wellcome-funded project 'Data Science Without Borders'[122], which brings together health scientists, data specialists, AI experts and policy analysts to work with institutes in Kenya, Ethiopia, Senegal and Cameroon.

Discussions are ongoing with a number of partners, including the Australian Research Data Commons (ARDC), the Helmholtz Metadata Collaboration, the Consortium of European Social Science Data Archives (CESSDA), the Korea Institute of Science and Technology Information (KISTI), the African Open Science Platform (AOSP), the Malaysian Open Science Platform, the LIFES Institute in Leiden, the CivicDataLab India, and others.

CODATA is seeking partners around the world for this initiative and invites collaboration to explore case studies to use and further refine and implement the WorldFAIR approach.

[121] Delipalta, A. (2024), https://doi.org/10.5281/zenodo.11110563
[122] https://codata.org/launch-of-the-data-science-without-borders-project/

APPENDIX: WORLDFAIR DELIVERABLES

Here are the WorldFAIR deliverables ordered by Work Package, with publication month and downloads as of the time of writing:

1. D1.3 'First policy brief', https://doi.org/10.5281/zenodo.7853170 - (April 2023) **609** downloads

2. D2.1 'FAIR Implementation Profiles (FIPs) in WorldFAIR: What Have We Learnt?' https://doi.org/10.5281/zenodo.7378109 - (Nov 2023) **1216** downloads
3. D2.2 'WorldFAIR's Experience with FIPs (second set of FAIR Implementation Profiles for each case study)', https://doi.org/10.5281/zenodo.11236094 - (May 2024) **35** downloads
4. D2.3 'Cross-Domain Interoperability Framework (CDIF) (Report Synthesising Recommendations for Disciplines and Cross-Disciplinary Research Areas), https://doi.org/10.5281/zenodo.11236871 - (May 2024) **99** downloads
5. D2.4 'Recommendations and framework for FAIR Assessment within (and across) disciplines', https://doi.org/10.5281/zenodo.11242737 - (May 2024) **31** downloads

6. D3.1 'Digital recommendations for Chemistry FAIR data policy and practice', https://doi.org/10.5281/zenodo.7887283 - (May 2023) **817** downloads
7. D3.2 'Training Package: FAIR Chemistry Cookbook', https://doi.org/10.5281/zenodo.10711950 - (Feb 2024) **174** downloads
8. D3.3 'Utility services for Chemistry Standards', https://doi.org/10.5281/zenodo.10514901 - (Jan 2024) **204** downloads

9. D4.1 'Nanomaterials domain-specific FAIRification mapping', https://doi.org/10.5281/zenodo.7887341 - (May 2023) **152** downloads
10. D4.2 'FAIRification of nanoinformatics tools and models recommendations', https://doi.org/10.5281/zenodo.10629631 - (Feb 2024) **80** downloads

11. D5.1 'Formalisation of OneGeochemistry', https://doi.org/10.5281/zenodo.7380947 - (Nov 2022) **216** downloads
12. D5.2 'Geochemistry Methodology and Outreach', https://doi.org/10.5281/zenodo.10406332 - (Feb 2024) **60** downloads
13. D5.3 'Guidelines for implementing Geochemistry FIPs', https://doi.org/10.5281/zenodo.10712808 - (Feb 2024) **81** downloads

14. D6.1 Cross-national Social Sciences survey FAIR implementation case studies, https://doi.org/10.5281/zenodo.7599652 - (Feb 2023) **661** downloads
15. D6.2 'Cross-national Social Sciences survey best practice guidelines', https://doi.org/10.5281/zenodo.8308012 - (Sept 2023) **199** downloads
16. D6.3 'Pilot Testing Harmonisation Workflows', https://doi.org/10.5281/zenodo.10724744 - (Feb 2024) **79** downloads

17. D7.1 'Population Health Data Implementation Guide', https://doi.org/10.5281/zenodo.7887385 - (May 2023) **180** downloads
18. D7.2 'Population health resource library and training package', https://doi.org/10.5281/zenodo.10010936 - (Nov 2023) **128** downloads
19. D7.3 'Population Health Data Policy and practice recommendations', https://doi.org/10.5281/zenodo.11242767 - (May 2024) **27** downloads

20. D8.1 'Urban Health Data - Guidelines and Recommendations', https://doi.org/10.5281/zenodo.7887523 - (May 2023) **400** downloads
21. D8.2 'Urban health data - learning and training', https://doi.org/10.5281/zenodo.10731625 - (March 2024) **65** downloads

22. D9.1 ‘Data standard for sharing ecological and environmental monitoring data documented for community review’, https://doi.org/10.5281/zenodo.7849241 - (May 2023) **396** downloads
23. D9.2 ‘Community consultation and finalisation of Biodiversity FAIR data model’, https://doi.org/10.5281/zenodo.10058058 - (Nov 2023) **76** downloads

24. D10.1 ‘Agriculture-related pollinator data standards use cases report’, https://doi.org/10.5281/zenodo.8356529 - (Sept 2023) **739** downloads
25. D10.2 ‘Agricultural Biodiversity Standards, Best Practices and Guidelines Recommendations’, https://doi.org/10.5281/zenodo.10666593 - (Feb 2024) **439** downloads
26. D10.2 ‘Agricultural Biodiversity Standards, Best Practices and Guidelines Recommendations: Tutorial’, https://doi.org/10.5281/zenodo.10688865 - (Feb 2024) **130** downloads
27. D10.3 ‘Agricultural biodiversity FAIR data assessment rubrics’, https://doi.org/10.5281/zenodo.10719265 - (Feb 2024) **253** downloads

28. D11.1 An assessment of the Ocean Data priority areas for development and implementation roadmap’, https://doi.org/10.5281/zenodo.7682399 - (March 2023) **436** downloads
29. D11.2 ‘New interoperability specifications and policy recommendations’, https://doi.org/10.5281/zenodo.10219933 - (Nov 2023) **207** downloads
30. D11.3 ‘Disaster Risk Reduction findings and recommendations’, https://doi.org/10.5281/zenodo.11074552 - (June 2024) **6** downloads

31. D12.1 ‘Disaster Risk Reduction Case Study report’, https://doi.org/10.5281/zenodo.7887557 - (May 2023) **141** downloads
32. D12.2 ‘Disaster Risk Reduction Domain-specific FAIR vocabularies’, https://doi.org/10.5281/zenodo.8110630 - (July 2023) **92** downloads
33. D12.3 ‘Disaster Risk Reduction findings and recommendations’, https://doi.org/10.5281/zenodo.11074552 - (April 2024) **23** downloads

34. D13.1 Cultural Heritage Mapping Report: Practices and policies supporting Cultural Heritage image sharing platforms’, https://doi.org/10.5281/zenodo.7659002 - (Feb 2023) **989** downloads
35. D13.2 ‘Cultural Heritage Image Sharing Recommendations Report’, https://doi.org/10.5281/zenodo.7897244 - (May 2023) **970** downloads
36. D13.3 ‘Implementation and Testing of Cultural Heritage Image Sharing Recommendations: DRI Case Study Report’, https://doi.org/10.5281/zenodo.10850009 - (March 2024) **160** downloads

37. D14.2 ‘WorldFAIR Sustainability and Exploitation Plans’, https://doi.org/10.5281/zenodo.11110563 - (May 2024) **30** downloads
38. D14.3 ‘WorldFAIR Outreach and Engagement Report’, https://doi.org/10.5281/zenodo.11205263 - (May 2024) **10** downloads
39. D14.4 ‘WorldFAIR Updated Dissemination, Communication and Exploitation Plan’, https://doi.org/10.5281/zenodo.11205161 - (May 2024) **4** downloads.

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No **101058393**