European Commission

Good practices and resources to improve the utility of research data in regulatory assessments

Webinar outcome report

Franco, A., Worth, A., Chinchio, E., Katsanou, E., Beronius, A., Agerstrand, M., Lynch, I., Rocca-Serra, P.

2024

Joint Research Centre

EU Science Hub
https://joint-research-centre.ec.europa.eu

JRC137088

Ispra: European Commission, 2024

How to cite this report: European Commission, Joint Research Centre, Franco, A., Worth, A., Chinchio, E., Katsanou, E., Beronius, A., Agerstrand, M., Lynch, I. and Rocca-Serra, P., *Good practices and resources to improve the utility of research data in regulatory assessments*, European Commission, Ispra, 2024, JRC137088.

Outcome Report Online Webinar: "Good practices and resources to improve the utility of research data in regulatory assessments"

Date: 31 January 2024

Organisers: EC Joint Reseearch Centre, Karolinska Institutet, Stockholm University

Background and objectives

Standard information requirements based on internationally accepted guideline methods, such as those adopted by the OECD, provide a trusted evidence base to chemical safety assessors. Chemicals legislation, however, requires assessors to consider all available scientific sources to identify relevant and reliable data feeding into regulatory assessments. This includes all kinds of data derived from scientific studies published in peer-reviewed scientific literature, curated databases and grey literature, which could inform hazard, exposure or risk assessments.

The application of modern methodologies in (eco)toxicology is resulting in a growing quantity of published, peer-reviewed non-standard (eco)toxicity data. However, standard information requirements are not evolving sufficiently quickly to embrace such developments. The structured approaches used in regulatory settings to assess study reliability and relevance are not routinely integrated within journal peer-review processes. Thus, assessors need to identify and evaluate an increasing amount of academic and non-standard data generated by a wide variety of different methods and models, with variable reliability and reporting standards. Compared to standard studies, assessing the reliability and relevance of non-standard research data is more challenging and time-consuming.

Over recent years, international organisations (e.g. OECD), national authorities, scientific societies (e.g. Society of Environmental Toxicology and Chemistry, SETAC) and communities of practice (e.g. Equator network, Elixir toxicology) have developed numerous guidance documents on good practice, reporting standards and tools for different types of scientific data (e.g. *in silico*, *in vitro*, *in vivo*, omics, human data). Regulatory authorities across countries and policy areas have developed guidance and workflows to aid assessors with the identification, screening and evaluation of academic and non-standard data for regulatory assessments.

The EU Chemicals Strategy for Sustainability stressed the need to improve the regulatory uptake of research data. Since this is a common challenge across countries and policy areas, an expert ground of the OECD Working Party on Hazard Assessment (WPHA) is currently developing a Guidance Document to define and promote good practices to improve regulatory uptake of non-standard scientific data. The guidance targets both the research community and regulatory assessors. The guidance defines good practices and general reliability criteria for regulatory consideration of research data. It will provide an entry point to identify existing resources, standards and tools available for specific data types.

The objectives of this webinar were:

- To raise awareness of the policy challenge and of the related OECD WPHA initiative
- To collect inputs from researchers active in the development of good practice, reporting standards or data management solutions

- To engage scientists in supporting the implementation of the guidance

Outcome summary

The webinar provided an overview of the state of play and the evidence base feeding into the development of the guidance. The first part consisted of four presentations, covering an introduction to the policy challenge and to the OECD project (Antonio Franco) and examples of recent and ongoing research activities aimed at improving tools and approaches for the identification and regulatory consideration of research data (Anna Beronius, Philippe Rocca Serra, Iseult Lynch):

- The challenge of using (non-standard) research data in regulatory assessments and the OECD initiative (Antonio Franco, JRC, OECD WPHA Expert Group on research data)
- *Experience and contributions from the SciRAP initiative* (Anna Beronius, Karolinska Institutet, OECD WPHA Expert Group on research data)
- *Enabling innovation: from data science research to regulatory application* (Philippe Rocca-Serra, University of Oxford, Molecular data production and management of PrecisionTox)
- *PARC perspective* (Iseult Lynch, University of Birmingham, co-lead of PARC WP7 on FAIR data)

The presentations session was followed by an open discussion (not recorded). The audience engaged on some of the crucial elements of the challenge, including:

- The long-term sustainability and accessibility of research databases and their interoperability with governmental databases.
- The utility to interconnect established research initiatives to governmental databases (e.g. future Common Data Platform on Chemicals to be hosted by ECHA, OECD eChemPortal) to ensure long term sustainability and workability. There is however no one size fits-all-solution for data findability and accessibility.
- The importance to provide access to the detailed procedure used to generate the data as well as to the raw data.
- The need to provide guidance by introducing and promoting good practice, available tools and approaches, while not to being prescriptive about the use of specific tools. This reflects the fact different tools may be preferable, depending on the jurisdiction and context, while sharing common basic principles. With this in mind, opportunities for harmonisation will be explored where there is consensus.
- The challenge to address evaluation of “relevance”, which is context dependent and changes over time. In this case, the utility of meta-data (e.g. AOP ontology) to support and harmonise relevance evaluations was highlighted.
- The importance to find ways for the two communities (researchers and safety assessors) to better communicate between them. Researchers are not aware of what is needed for their research to be used in regulatory assessments.
- The utility to publish datasets as final outcomes of scientific projects. Papers presenting the final dataset at the end of a project are very useful and also usually well cited.

Transcripts of presentations

Antonio Franco (European Commission, JRC)
This is the agenda of the webinar. I'll start with a brief introduction on the challenge of using non- standard research data in regulatory assessments and the related OECD initiative.
We'll then have Anna Beronius who will share her experience and wisdom on the development and use of the SciRAP tool. She's also representing like me the OECD, WPHA expert group on Research data.
Then we'll hand over to Philippe Rocca Serra who will explain how Precision Tox is coping with the facilitating and supporting the transfer, the translation of research data from the research field to the regulatory application.
Last, but not least we will have Iseult Lynch. Her presentation will give us the PARC perspective on this challenge, especially in relation with her role of colleague of the work package on fair data. The big work package on fair data of the PARC partnership Then we'll have the Q&A.
So, what are we talking about is research data and I guess some of you are wondering whether this is the same initiative, the same project as the one you heard some time ago, which we used to refer to as the academic data project and initiative, and yes, it's the same thing.
The new terminology we have adopted is this of research data because that reflects the latest definition that defines the scope of this initiative. So, research data is defined as any data that can inform hazard exposure or risk assessment regulatory assessments.
It is generated by scientists from academia, but also from public and private research institutes industry or NGOs and it's data that is published in peer reviewed, scientific literature most typically, but also in curated the databases or grey literature and typically. This is where we are mostly focusing on as part of this initiative. This data is not carried out to inform assessments regulatory assessments and it's typically generated using non- standard non- guideline experimental or computational methods. The research data can inform different steps from data requirements to hazard assessment, exposure assessments or risk assessment. So why do we bother with this? Why do we need to use research data in regulatory assessments?
The first reason is from a scientific perspective to the problem. Research data can add valuable information. It can add information that is not available through standardized regulatory type of study data. This could be, for example, in the form of additional endpoints in existing standardized studies or coming from study types that are not covered by internationally accept test guidelines. For example, mechanistic data from all sorts of modern toxicology tools. Epidemiological studies are another prominent example.
But there's also a legal obligation to do this because virtually all pieces of chemical legislation that are in scope here have some sort of provisions that require assessors, both regulators and registrants, to consider all available scientific data in performing assessments.
And this is where this initiative is coming from. This provision is difficult to implement because the research data domain is huge and there's no hard coded rules to define what is considered relevant and reliable for regulatory assessment.
This puts a heavy burden to assessors and especially to regulatory agencies. And we have heard from the experiences of the US EPA, from Canada from EFSA that, where hard coded or anyway well-structured approaches are implemented to make sure that all relevant and reliable scientific data is screened and considered, this takes a very big effort. So, there's a major efficiency challenge to this problem, but there's also a regulatory coherence challenge to the problem because the way we go about this, at least in the EU is not the same between agencies and the scientific committees. The different policies domains have developed their own let's say solution to tackle this this challenge.

So, as I said, the initiative originated from the chemical strategy. We organized a workshop in 2022. That initial work led to the establishment of an expert group at the OECD and of a WPHA project with the aim to develop an OECD guidance. As part of that we completed two surveys. We are currently running four case studies to explore good practice on how to best use research data to address regulatory problems. And we will aim at publishing the guidance by the end of this year. This webinar is part of this process. It feeds into the development of the OECD guidance and promotes the implementation of the guidance.
I'd say a few words about how this is looking like. This is our graphical representation of the project. As you see there, the way we are looking at it is to consider the life cycle of research data that goes from data generation to documentation, reporting, retrieval of data for evaluation and use in regulatory assessments. Several stakeholders are involved along this process, and we are doing our best to make sure that all perspectives are considered in the formulation of this guidance document.
The guidance is structured in three parts. The first one gives guidance on the production and reporting of non- standard data for regulatory considerations and targets primarily the research community. This is where the focus for today is. There will be then a section about principles and approaches for the identification selection and evaluation of research data. Sorry, that term is not yet updated in this slide. So that's the bottom half of that cake graph. Finally, we will provide some recommendations regarding opportunities for harmonization of data evaluation extraction and reporting.
And with that, I just remind on the three objectives that we set for this webinar: we want to raise awareness on this policy challenge and the OECD initiative; but very importantly we want to collect inputs from you researchers, especially if you're active in the development of good practice reporting standards, and data management solutions that facilitate regulatory uptake. And we want to engage the scientific community in supporting the implementation of the guidance, once this is out. With that, I will welcome Anna to take the floor for her presentation.

Anna Beronius (Karolinska Institutet)
Thank you very much. Antonio, I hope that you can still hear me. Okay.
I will talk a little bit about the SciRAP or Science in Risk Assessment and Policy initiative and share some reflections maybe and experiences from that. I also present it as an example of a tool that can be used to help both researchers and assessors in this goal of improving the use or utility of research data in regulatory assessments.
So, the Science in Risk Assessment and Policy, or SciRAP as we call it, is a research initiative. You see the, the core members of this group. There are also other colleagues that worked on this, but the core group is down here. It's us from Karolinska Institutet together with our colleagues at Stockholm University, we all do research in different areas or regulatory toxicology, mainly focusing on developing and furthering methods for hazard and risk assessment, investigating processes for risk management and so on.
In this work, started about a decade ago, we identified this need for tools to improve structure and transparency in how toxicity and ecotoxicity data are evaluated for hazard and risk assessment and we also identified a need to bridge the gap between academic research and regulatory assessment. So, basically with this initiative and project we wanted to contribute with a user-friendly tool that could facilitate structured data evaluation. I'm only going to have time to really introduce the tool today. There's much more available online, if you're interested. I put some selected publications here, but there's more on the SciRAP website and of course, I'm happy to answer any questions.
The SciRAP tools are openly available online at scirap.org. Anyone can use it and you can go and explore it in your own time. We've developed tools for study evaluation and in parallel with that also guidance and checklists for reporting. The tools for study evaluation are mainly intended for assessors to facilitate structured and transparent evaluation of reliability and

relevance of ecotoxicity and toxicity data. The reporting checklists are mainly intended for researchers, who want or need some assistance in reporting data in a way that maximizes the use for regulatory assessments. So currently we have tools for ecotoxicity data, also called the CRED criteria and there's a version of this for nanomaterials, called nano-CRED. We have tools for the evaluation and reporting of in vivo animal toxicity data, and in vitro toxicity data as well as specifically for nanomaterials. And we're currently working on developing a tool for evaluating observational epidemiological studies.

The SciRAP tools are criteria-based tools, which means that we've developed specific criteria for evaluating reliability and relevance and these are all based on requirements and recommendations in current OECD test guidelines, as relevant for these different study types. When we talk about toxicity studies, we further divide reliability into evaluating reporting quality and methodological quality. So, if you're not familiar with this, the reporting quality is really about the completeness of the reporting of the study design conduct and results, whereas methodological quality is about evaluating the appropriateness of the study design and conduct, including the sensitivity of the model and validation, repeatability and so on. Relevance is the extent to which the study or data set contributes with appropriate information to answer a specific problem formulation or assessment question. So, again these tools are intended for assessors, but are also quite useful for researchers, who want to familiarize themselves with which aspects assessors look at when they evaluate study reliability and relevance in regulatory assessments.

So very quickly the output of the SciRAP tool is mainly qualitative. It's this colour profile, and again, I don't really have time to go into it in detail, but you also have this more detailed evaluation or detailed summary of the evaluation of the study. This panel is showing a summary of an evaluation of methodological quality of an in vitro study. So, when you export the SciRAP results you get this. Here for the tox tools you also get a score, which is a numerical score, but the focus of the output should be on this colour profile.

I just wanted to show some how the SciRAP tools can be used, with just a few examples. This is a study that we conducted where we looked at, studies that had been included in REACH and we re-evaluated them. They are usually categorized according to the Klimisch categories as reliable without restriction, reliable with restriction, not reliable and not assignable. And we setup as system based on the SciRAP tool to do this. You can just see here as an example of how the SciRAP tool can be used. One thing with the SciRAP tool is that it has these criteria that you evaluate, but the output is not automatically a categorization of reliability or relevance. So, these are principles that you have to set up on a case-by-case basis based on the purpose of the assessment that you are conducting. So, this just shows, for example, how the, this quality assessment can be then translated into these reliability categories, the Klimisch categories. So, for example, for a study to be reliable without restrictions it should be well designed and performed, all key reporting and methodology criteria are judged as fulfilled and, there are no deficiencies in other non-key criteria that are considered to affect the reliability of the study or make the study not assignable. These are principles that we set up on a case-by-case basis.

Another example is the assessment for endocrine disrupting properties of bisphenol-F, where we did a similar principle but set it up a little bit differently also taking the score into account. But I really want to point out that the qualitative output of what we are looking at. So, just some examples for someone who wants to go and have a look at how this tool can be used.

When we look at the ED assessment, you also, of course, you also have to look at the assessment for complete lines of evidence and do a weight of evidence assessment and this just a chosen example, then of how we use this tool to do the line of evidence assessment and weight of evidence assessment to evaluate lines of evidence as strong moderate and weak. These provide examples of how you can use a tool such as the SciRAP tool to in a very structured and transparent way to include non-standard or research data into assessments.

Here is another example not from us, but from another group. This colour profile is used to visualize reliability and relevance across the studies in a line of evidence. This provides an example of how others have used the tool and really applied this idea of using the colour profile, this quality assessment.
So, as I said, the tools themselves, the assessment tools can also be useful for researchers in trying to understand or understanding increasing their awareness of which aspects regulators look at when they evaluate study reliability, but we also have these SciRAP reporting checklists, which are also available online and if you go to the ScirAP page, you see them listed like this. This is, of course, longer, it continues down here. You can also download it as an Excel template where you can fill in your data and even submit that a supplemental material, if needed. This is an aid for researchers to facilitate also evaluation for the assessors when all important information is included in a structured way. And it promotes transparency.
Finally, I just want to share some reflections from our work with the SciRAP initiative and other projects as well. I think it's important to remember that academic research on chemicals and regulation are two different spheres. We can certainly benefit, or we want to benefit from the overlap between these spheres. But they are still separate and there is a need also for basic research for the sake of research. So, I think it's important to remember that within the academic community. We have many researchers who are perhaps not aware of regulatory requirements and that we need to have still this flexibility. It's not about forcing all academic research into one standardized format, but it's how we can make better use. And for this, I think we need to think of this as a two- way street. We need to support both researchers and regulators in improving the use of research data in regulatory assessments.
For researchers, as I said, we may need to improve awareness or increase awareness in the research community, but we also need to provide incentives. We have to remember that academic research is driven by novelty by exploring new things, not by repeating or validating necessarily results that have already been made. We don't always have the possibility to do this either. We're not funded, not giving given ethical permits to repeat, for example animal studies. So, incentives in the form of funding, and in the form of giving more credit, so to speak, to conducting this type of research is needed. We also need tools and guidance, so that it's clear what type of information is needed and in which format.
And I think we can really benefit from having positive examples from the regulatory arena, because we know that research data are being used for regulatory assessments, for example, for restrictions under REACH, or in recent EFSA assessments. These positive examples, I think, can really provide a good leverage for communication. Similarly, then for regulatory assessors, I think we need to provide tools to increase familiarity and acceptance of non-standard data. And we need tools and guidance, such as the SciRAP tool or other tools, that can be used to facilitate structured and transparent evaluation.
And finally, I think education and training of course is very important to tackle all these challenges. That was my short presentation. I of course want to thank everyone who's involved in working with SciRAP, and thank you for your attention.

Antonio Franco
Thank you very much Anna.
As I said before, we can take questions just after the four presentations. With that, I would like to welcome Philippe. While he puts up his presentation, I just say that he's one of the pioneers of the FAIR data initiative and he represents here Precision Tox and leading the working group on molecular data production and management of that project. So, Philippe, we're looking forward to this.

Philippe Rocca-Serra (University of Oxford)
Hello, yes, good afternoon.

Thank you for the kind introduction, way too kind.
But I'd like, yes, indeed, to build on the presentation by Anna Beronius and insist about provide a spin on what we have done as part of the ASPIS cluster with this kind of more fundamental research and more omic type of data.
This is the key aspect that we'd like to insist and present during this presentation. I start the presentation with presenting to you what the ASPIS cluster is. It's basically three projects funded by the European Union brought together. They are all operating in the same space, that of chemical exposure and safety assessment, but each looking at different angles, complementary angles, and this is why the ASPIS cluster has been created to synergize and converge a number of initiatives so that the outcome of all these projects could be maximized for future re-use. And this is in a way where the notion of FAIR will be relevant as I will expose a bit later.
So with this, I think going back to the common point of all these three projects where we look at the effect of exposure to chemicals at the molecular level. We explore the effect of these xenobiotics in biological system by means of massively parallel methodologies, molecular endpoints that are generated used using modern technologies, such as next generation sequencing, mass spectrometry or mass spectroscopy. The idea is to characterize the molecular responses.
The project uses an array of organisms ranging from cell lines or cell cultures, also reusing existing data. But in the case of Precision Tox, we're also looking at non sentient organism. And these are the tests on species that are used in the consortium as part of a new testing paradigm to do assessments by tapping into the aspects of philotoxicity. The notion of this is really the main angle that we have.
The aim is to understand the activation of toxic pathways by developing new approach methodologies. Most of them will be building on the combination of obviously modern method, analytical method from artificial intelligence and the tooling that is available around that, but also taking advantage of new data structure for representing information in a form of ontology knowledge graph for instance. This aspect is important in order to relate the classic endpoints, which are used in a chemical assessment safety assessment, with the omics endpoints and to combine more endpoints to improve the capability of the detecting the point of toxicological departure, but also being able to do grouping of chemicals to perform safety assessments at scale, bearing in mind that we cannot test all possible chemicals. The evidence needed to build efficient and reliable read across techniques is key.
So why does FAIR matter in this context? FAIR is in short, stands for findable, accessible, interoperable and reusable. Essentially it means that we have self- describing data that can be understood by machines, by software agents, and this is a notion of machine activity. We need to be able to access resources that that software agents can understand, and this means that we have.

Structure where the semantics has been clarified. This is why I refer to explicit semantics. And the notion of FAIR really allows to protect the investment by the EU Commission. For instance, in the provision of these datasets simply by ensuring ruse of the data at scale. This is done through ensuring that all the research data is always shipping with sufficient metadata that enables understanding of the study design any confounding factor that could be associated with the data. So, if we think about using fair to organize the data collection, we need to embed that into the practice from the onset. So, it starts with the building, a good data management plan and this can start with a survey of the art to understand the resources in terms of vocabulary, data structure, formats, syntax that are available to the domain. To promote reuse there are resources such as fair-sharing that are there to help you. We don't need more standards. I think it's more a matter of converging on agreeing on a set of, of resources that should be used. We should agree in terms of information requirement what should be reported for the particular

data model and which vocabulary do we need to use to annotate those elements. This is to prepare for structuring the information from the lab or the robotic platform for that acquisition all the way down to the representation as knowledge graph. The evidence gathered using all these technologies requires having this upstream work done before hitting the lab.
The next aspect is that in a project such as precision talks, we really try these principles from the onset and we were lucky because we had this pilot phase, which allowed us to test a bit some of these hypotheses, and to calibrate our understanding.
The pilot phase was a perfect situation for us to address the issue of phasing the projects. We are about to collect the data, but the tools are not ready, so we need to stop measures. Most of the time, at least a very lean process to start collecting the initial data is usually done in the form of spreadsheet templates with naming conventions embedded. But what we wanted to do is bring the FAIR practice even more upstream, starting from the plan to initiate a data collection. We use the pilot project in Precision Tox to refine the user requirements and the annotation requirements talking to the various subject matter experts, develop a number of software prototype do user testing and then refine the related production version. All along we are using software engineering best practice to ensure that we have test driven development, continuous testing documentation both for our users and us. All of this belongs to the movement of FAIR because it's all the digital objects that are manipulated that should be handled with such similar care.
In the next few slides, I will rapidly show you the kind of approach that we've done and how we build meta data manager, which uses the information about the study design to guide the creation of the metadata template that we all have. It's powered by a data model that we are maintaining here in Oxford: the ISO model with a powerful API. But we wanted to harness it not to use it after the fact, after the data acquisition, but very early on, in the process at the planning phase so that we could deliver to the people in the lab, the templates that could help them carry out the experiments.
And we also tested a number of connections with robotic platforms to be able to capture the kind of package effects or in the case of spectrometry, for instance, the kind of quality control that should be also reported to ensure that down the line people can access the raw data and re-analyse everything, should there be a need.
So, rapidly I will go quite quickly through few slides just to highlight that we handle everything in open source and using the GitLab infrastructure for all our software development. But the key point is that we moved all the practice very upstream pre-lab. We are in a position where we can generate all the metadata annotation framework and templates before people go into lab.
And they can use this on an app in the lab or as a spreadsheet, which is connected to the annotation and looks up on the identifier to the database. This information is then persisted to the database, which can be also looked up and all samples are identified.
This is a few screenshots about the app itself, highlighting the importance of the study design and linking to the chemical information that we have also stored, linked with identifiers from the relevant repositories we've defined. Here we are just showing you an output of the spreadsheet with a naming convention, which is a short tag that is meaningful for people in the lab operationally as well. It is not a long string, but it's kind of a hash identifier as well as the machine-readable format that we have as it is.
This is for precision talks, but now I'm presenting also a similar approach that has been used by Risk Hunt3r. They also used the spreadsheet templates obviously to collect the data produced at different sites. And they deposit the information to bio- studies at the EMBL-EBI. I will get back to that point in a moment. Also, everything from the metadata is then uploaded in the context of RiskHunt3r to the BioStudies repositories and database. Then the information is available through REST API that can be accessed through a number of means either R or

notebooks or Colab, which can be executed as well. That's a way of enhancing machine readability.
This is a pointer to the actual resource that hosts the results of the data collected by Risk Hunt3r. I encourage to you to contact Barry Hardy at the project for more information.
ONTOX also - this is a very busy slide, but I highlighted in red the boxes. Convergence is the keyword here where, again the common point with a ToxTemp SOPs are used to collect the data into the ONTOX project, shared with Risk Hunt3r. The data is persisted to the bio studies and then the knowledge graph is generated in the form of the new ASPIS4J endpoint.
Going back to Precision Tox I think it's important to see that aspect of convergence. We've been collaborating with Elixir as well to be able to support the deposition and the publication of omic data through the public repositories, the institutional repositories as well. This is was supported by two projects. One looked at the brokering aspect. Because there are no multi-omic repository that exist, we need to dispatch the information to the relevant repositories. The other project was about how we provide a machine-readable object that contains both the raw data, the metadata, the results of analysis as well as the computational workflow in one single object that could be accessed and mined by machines.
This is in a way two objectives here. As I started with my presentation highlighting the fair-sharing repository where is out starting point to start a data management plan. I think this is to promote reuse of the format and the existing resources already. This is tested both in the fair cookbook resources that has been produced a few years back now for guiding how to implement that in real life, but also pointing you to recent publication in GIGA Science, which uses the research object as well to package everything on experiments.
This is the last screenshot of a kind of overview, where you have the data sets itself, which is formatted with a standard metadata model. The raw data is available as well, but the research object package everything in one single resource, including the computational workflows, which are deposited to Workflow Hub, which gives you a unique identifier, which we can use as a citation. But also, you can execute some of the workflow that have been used to analyse the data and check yourself if the result that I claimed are backed up by the computation. So, this is the notion of trust and I wanted to insist on this because to establish research data for the context of regulatory assessment, I think this is the kind of starting point that we need to establish: the ability to access the data and trust it.
The last slide that I have is simply to highlight the need as well to keep in mind the notion of how we represent the conclusion and knowledge that we have gathered in new data structures, such as this graph owing to their relevance. First one, two things, the ability to connect knowledge bases using stable identifiers, but also their importance for machine learning approaches. And this is also something I wanted to highlight.
I will stop here. I'm already over time, I realize. Thank you very much.

Antonio Franco
Thank you, Philip. Thank you so much for your very insightful presentation. I'm sure we will get back to that during the discussion session. Iseult, over to you. Iseult is work package lead for the FAIR data work package of PARC. PARC, you know, is a very big partnership. I just checked this: two hundred partners. Each of them producing data, I suppose. And Iseult is responsible to make them all as FAIR as possible. Tell us how you do it.

Iseult Lynch (University of Birmingham)
Yes, exactly. Thank you very much. It's an honour to be here and it's fantastic to see how much convergence there is already in the PARC approaches with the ASPIS approaches and also, I'm delighted to say that we have some activities, some joint activities coming up in the near future for further that harmonization and integration. I will give a little bit of a whistle stop tour to some of the approaches that were taking in PARC, which is the partnership for the assessment

of the risks from chemicals. It's a seven-year project. So, we are now just past the first 18 months, coming up to 2 years in April. Time flies. In that period, we've really, I guess, spent a lot of time training within work package 7 on the technical aspects of FAIR. And now we are at a point where just actually this Monday we had fifteen PARC data, GoFAIR fellows inducted. So that's a, a big milestone and a lot of work went into that. PARC then as a project, there are nine work packages overall, but four of them are the key data generating work packages that I'm going to present.

We have an entire work package dedicated to human biomonitoring, which builds in large part from the previous Human biomonitoring for EU consortium and then also looking at chemicals in environment and food products, etcetera.

Then work package 5 is looking at hazard assessment, really focusing on the new approach methodologies, and how we drive those to a point where the data can be accepted and utilized in regulation. Then we have work package 6, which is looking at risk assessments. So, pulling the

Exposure and hazard together to look at overall risk assessment using next generation approaches and integrated approaches to testing and assessment and so on. All of that should come together in the PARC toolboxes, which is work package 8. They will cover safe and sustainable by design, early warning systems, integrated modelling and so forth.

Then work package 7 sort of sits in the middle here and is managing all of those data flows, supporting users in making their data FAIR. In work package 7, which is the fair data work package. We have sort of three major core activities that mapped those data generating work packages. So, we have a large activity around everything we need to do to make human biomonitoring data FAIR, what we need to do to make environmental data FAIR and what we need to do to make toxicology data including omics workflows FAIR.

There's a link there to the PARC, “A walk in the PARC” publication, which sort of lays out the project overall. We have also been doing a lot of work on sort of trying to set what our ambition should be for PARC and how we do that, how ambitious we should go or whether we should aim to have everything a little bit FAIR or something fully FAIR and linked, or somewhere in the middle.

We've taken the approach in our PARC data policy that we will use an approach that defines data from re-useless, so you can't reuse it again to data that is at least findable by applying persistent identifiers, so forth as we go up where the data is FAIR and can be either open or closed depending, as with some of the human biomonitoring data, if there will be personal data in there. This will be protected data that you won't be able to be open, but it can still be FAIR. We can still know the metadata. We can still know how you can get it and what the access conditions are. This is where it’s FAIR and open. And then the, the absolute target would be if it's fully linked as well.

We'll strive for a minimum level of FAIRness for all PARC data and metadata, whether it's open or closed, and that would be the equivalent of the D and E. Our target, our key performance indicator is that 80% of PARC data will achieve that. Now it is a cumulative score that we have. So that will be by 2029 rather than 790% each year or even 80% percent. So, it will, it will start low and go off.

Then we also had quite a bit of discussion around what sort of data we're talking about. Are we talking about only the summary data, only the metadata or do we also need to be able to access back to processed data and the raw data. We will be making recommendations and approaches for all of those, but the minimum will be to have summary data and all the metadata FAIR. We have been working in PARC quite closely with the GoFAIR foundation. We've spent a lot of time training on the GoFAIR approaches. The FIP is the FAIR implementation profiles. They are a tool that I think will allow us to address some of the issues that Philippe raised about the availability of existing standards and just the need to build consensus. So again, utilizing a FAIR implementation profile you can define your research community, and then define or

declare what FAIR enabling resources your community is going to use. And in some of those, they are technical and in some of them they are community standards. They may be at various levels of maturity, so they may be things that you're currently using now but will be replaced when something better comes along. Or they may be something that isn't yet developed but is aspirational. But by having those declarations of your FAIR implementation profile allows the next project or the next group not to start from zero. They can go: Okay, here's the decisions that were made and the rationale for those and a sort of a time scale for them.
Our work at the minute is to have a reference FAIR implementation profile for all of PARC. And then some domain specific ones or data type specific ones as well. And that's where we have the community, the domain experts really telling what is needed in addition to the metadata that would be PARC level.
We're also working on a number of PARC ontologies and PARC vocabularies, aiming not to start from scratch, but to build and update and amend as needed existing ones, and similarly building the, the PARC metadata schema. So again, very similar approaches to what Philippe presented for Precision Tox and ASPIS. The fact that we're all converging is also really nice. We'll have good interoperability.
As I said, we're working closely with the GoFAIR Foundation. What is helpful here is that the FAIR Foundation have differentiated between some of the red principles of FAIR that are the technical ones, and then the blue ones that are sort of a community consensus. We've been building our FAIR implementation profiles. We are now working to enable those FAIR implementation profiles to feed directly into our project level data management plans. So, and within the context of PARC as a partnership, within that we have around a hundred or so individual projects that are running. That will enable us then to ensure that we know what data sets are coming and how each data set is being managed where it's going and so forth and facilitate linked data. I will come back to the issue of data sets in a minute because that's one of our current challenges.
The FAIR implementation profile is socio technical. Machine actionable are all the ones in yellow and that basically means that it is relying on humans to do things. The more we can make machine actionable the more it will facilitate regulators and researchers to find things. The red ones then are the technical ones. This is where we can make a decision on behalf of our PARC community, as to which repository we deposit the data in, which persistent identifier we are using. So, whether it's for chemicals, whether it's the INCHI or SMILES or so forth. The social ones, this is where we're spending our time in PARC. These are the ones where we have to determine what the metadata is and how rich that needs to be, what vocabularies we're going to follow and so forth. All of them feed into our declaration of all the FAIR enabling resources that this community will be utilizing.
Building on that then we'll have to PARC data hub. This is the platform that will provide integration between all the different types of data in PARC, all the different stakeholders of PARC and integrators with the broader community. We have identified a number of uses for the PARC data hub: finding resources related to chemical risk assessment, finding resources to help make your data FAIR, how to publish or deposit your data, accessing data and exchanging data with regulatory partners. We have begun to work out the steps involved in each of these. For example, to make your data FAIR, the user is the data owner and there are various steps being worked through with the guidance to generate a FAIR data package. Similarly, as a data publisher you want them to be able to exchange with IPCHEM or with the OECD and IUCLID and so forth.
I'm not going to go into the detail of the workflow, but an important piece is that we need these solutions to be accessible to experimentalists as well. One of the barriers that we've encountered numerous times is that we are in our work package, developing all our lovely technical solutions and our experimentalists are going “Yes, but I use an Excel sheet. That's what I want to do”. We have to then bridge those gaps and make sure that in the end, we can

come up with a template that is in an Excel format or a tabular delineated format that the researchers can download. We're using an approach called a Bag of Variables or the catalogue of variables, and this is just what it sounds like: the whole range of things that we need to have descriptors for, including chemical identifiers, biomarkers, also questionnaire and qualitative data and lab results at different scales and different degrees of complexity of data. That bag of variables approach is something that we're using across the work.

This is just another of the workflows. This is, for example, the data publisher and how to publish the bare data package. I'm not going to go through the details again, partly because I don't fully understand them. But the key point here is that some of them will be domain independent. Some of it will be just having your dataset available, for example in Zenodo. Some will be more domain specific. For example, IPCHEM for the human biomonitoring, which will of course in due course end up in the EU open data portal. Also, a lot of work has been done on toxicology data with Elixir and the Elixir toxicology community, and of course, ASPIS. Similarly on the environment data, a lot of work has been done in Norman and we're collaborating very closely with the Norman network of databases and in due course also with the IRENE research infrastructures. So, we know we are not doing this in a vacuum, but we also have the advantage of being a seven- year project, so we have quite some continuity. We can pick up best practice from others and help to build on it and disseminate it.

Other things that we will do, and I really don't have time to talk much about these, but I just want to highlight them so that people can get in touch if they want more information. We are also developing a lot of tools for data enrichment, so for text mining, for uncertainty analysis, for integration of meta-analysis and so on. So quite a bit there including mining of adverse outcome pathways and things. We are also building on various FAIR maturity indicators that have been developed. As we identify repositories that PARC partners are putting their data in, we are also running some of these assessments to see whether, for example, they are meeting the FAIR indicators of whether additional work might be needed to further increase the fairness of data in those repositories. This is a critical piece of work because it'll come back to our key performance indicators of how many of our data are FAIR. We've also been talking a little bit about a PARC protocols repository and our link to protocols.io, for example, to ensure that each of our studies is underpinned by the relevant protocol.

Of course, the APIs and the interoperability layer will all be critical, as well as integration and automation of various tools, like the SciRAP that we heard about and the ToxR tool, so that we can really get scores for completeness and leveraging the rich metadata also maybe for quality of the dataset.

One of our challenges is, how we define a dataset, though. We have as our key performance indicator, the number and the overall percentage of PARC datasets that have been made FAIR. As I said, our target is seventy to eighty percent by 2029. But we are of course stumbling at the first question of what a data set is. Is it the overall investigation, is it the study and its associated assays or is it at the individual assay level? Within that, then our investigation might be the equivalent of what we call a project in PARC. But even a project in PARC may have multiple investigations. So, we're still teasing out a little bit what is the unit of a dataset that we want to measure.

For the moment we are taking a somewhat simpler approach. We're using the unit of a publication and the data set underpinning publication for our initial evaluation. That will be useful because we're already identifying several simple things that PARC partners can do to make their datasets already more findable and more FAIR as they're publishing them. We'll be putting together, or we have already put together a checklist for people as they're publishing. Then we are coming up also with our FAIR score. That is how we then come up with the visualization of our achievement of our KPI. So, the data steward wizard, which is the data management plan platform that we're using, has some indicators that can give sort of percentages of findability. Also, the FAIR metrics for databases that I mentioned will be a key

part and all of the databases that we use will also have a FAIR implementation profile associated with them, and we will have those then listed as approved PARC repositories in our reference fair implementation profile.
That was really a whistle stop tour. But I do see an awful lot of synergies as what we've already heard in the previous presentations. I think the OECD initiative is really timely and fantastic and yeah, looking forward to contributing further to it. I shall stop sharing now.

Antonio Franco
Thanks Iseult. Very well delivered. And while we see all the virtual round of applause, let me just thank again all the speakers. I do feel a bit guilty to ask you all to summarize in ten minutes such cross- cutting broad challenges. We run slightly over time, but I think you all did a great job.

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:
— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
— at the following standard number: +32 22999696,
— via the following form: european-union.europa.eu/contact-eu/write-us_en.

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

Open data from the EU

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society

EU Science Hub
joint-research-centre.ec.europa.eu