Plan Overview

A Data Management Plan created using DMPonline

Title: improving Reproducibility In SciencE (iRISE)

Creator: Rachel Heyard

Contributor: Rachel Heyard, Gillian Currie

Affiliation: Other

Funder: European Commission

Template: Horizon Europe Template

Project abstract:

Structured understanding of the drivers of irreproducibility and presenting concrete solutions of tools and interventions will help to increase the quality, reliability and re-usability of scientific evidence. To this end, iRISE proposes to provide theoretical and empirical evidence of the effectiveness of specific interventions, and a framework for a robust, evidence-based road map for the development, assessment and implementation of interventions intended to improve reproducibility. iRISE brings together qualitative and quantitative expertise, from academia and SMEs, including meta-science, statistics, economics, artificial intelligence, research ethics and integrity, quality assurance, and project management. iRISE proposes the development of a general framework for diagnosing and addressing reproducibility problems using analytical and computational modelling, simulations and meta-studies. Data on existing interventions will be systematically curated and evaluated, and stakeholders will be consulted to collaboratively identify practices and tools that should be prioritised for implementation. iRISE proposes to conduct empirical studies of both technical and practice-based solutions to increase reproducibility. Across all iRISE activities, the influences of research culture will be investigated, with a focus on mainstreaming systematic integration of equity, diversity and inclusion practices. A comprehensive Stakeholder Forum will be engaged to provide advice, and iRISE will commit to open and reproducible practices. The different types of evidence generated will be integrated into an open knowledge base to support the community in decision-making to identify, test, and implement effective and feasible solutions for reproducibility. The members of iRISE have made pivotal scientific and policy contributions relating to robustness, rigour and reproducibility in the past and have the skills and tools to succeed in this ambitious project that has potential scientific, economic and societal gains both in Europe and beyond.

ID: 140490

Start date: 01-09-2023

End date: 31-08-2026

Last modified: 28-08-2025

Grant number / URL: 101094853

Copyright information:

improving Reproducibility In SciencE (iRISE)

Data Summary

Will you re-use any existing data and what will you re-use it for?

Almost all iRISE work packages (WPs) will re-use data, which is either publicly available, generated by an iRISE partner in previous projects or provided to iRISE via its cross-consortia collaboration with the two other projects funded through the same Horizon Europe Call (OSIRIS and TIER2). Specifically,

- **WP1** will re-use publicly available data on studies assessing the reproducibility of a finding or a field of studies, such as the FORRT (Framework for Open and Reproducible Research Training) Replication Database available via forrt.org/apps/fred_explorer.html), other reproducibility projects, as well data from a meta-analysis including animal and human studies, to study translation. The re-use of this data is part of Tasks 1.3 and 1.4 where metrics and theories will be tested on real data.
- **WP2** will re-use the SOLES R-package maintained by the CAMARADES group in Edinburgh, which includes a collection of tools. The R-package will be used to efficiently implement the iRISE Systematic Online Living Evidence Summary (SOLES). The SOLES R-package itself re-uses tools by creators outside of iRISE and the Edinburgh group who will be credited transparently.
  WP2 will re-use a dataset populated by the consortia OSIRIS (osiris4r.eu) and TIER2 (tier2-project.eu) during their scoping reviews on interventions to improve reproducibility. This dataset will be (re-)used to train the SOLES.
- **WP3** will re-use data provided by the OSIRIS and TIER2 consortia when gathering information for a best-practice guideline.
- **WP5** re-used the IICARus study protocol's (osf.io/xpu9p) structure to design the template for the unified study protocols for the iRISE intervention studies (task 5.1).
  Task 5.2 will use a series of available automated screening tools that were previously developed by one of the iRISE partners. The tools will be applied to a collection of manuscripts received from a publisher.
  Task 5.3 will use data and computational scripts that have been submitted by authors to run a computational reproducibility review.
  Task 5.4 will re-use questions from a public and validated survey on research climate and/or culture. The selected questions will be outlined in the study protocol.
  Task 5.5 will use publicly available data from biomedical research to inform its simulations. The task will also re-use code scripts previously developed by the task-leading partner.
  Task 5.7, a new task added, also known as "Table 1 project", will re-use bibliometric data to test and validate the pipeline.
- **WP6** will summarise existing and publicly available data from intervention studies for the framework developed in task 6.3. Task 6.6 will bring together existing (publicly available) training resources to develop train-the-trainer courses.

What types and formats of data will the project generate or re-use?

The iRISE consortium will generate and re-use various types and formats of data. Note that our definition of data is a broad one. In the following, *data is defined as all information or material needed*

to reproduce iRISE's results and outputs. This includes datasets, software, code, protocols, workflows, tables, images, videos, transcripts, articles.

- iRISE will comply with best practices and pre-register the **protocol** of all planned analyses whenever possible as text in PDFs or HTML on the Open Science Framework (OSF). The following tasks will produce at least one pre-registered study or analysis protocol: 1.1, 1.2, 1.3, 1.4, 1.5, 2.1, 2.2, 3.2, 3.3, 3.5, 3.6, 4.1, 4.2, 5.2, 5.3, 5.4, 5.5, 5.7.
- Tasks 1.1 and 2.2 will generate **literature review data**. The file format of these data outputs will be RIS and annotated CSV, specifically for data extraction. iRISE literature libraries used may also be shared via Zotero libraries.
- Some tasks, specifically 1.3, 1.4, 2.1, 2.2, 3.6, 5.2, 5.3, 5.4, 5.5, 5.7, will generate **computational or statistical analyses and code scripts** . These will be prepared in open-source software (R and Python). The iRISE SOLES (task 2.1) will be hosted as, and interacted with, through an RShiny app, which is based on R, CSS and HTML files (all non-proprietary). iRISE SOLES uses large language models and all prompts will be stored as Jupyter notebooks (.pynb).
- **Texts, data and metadata extracted from published papers** will be used in tasks 5.2 (PDF and XML version of publications with metadata), task 5.3 (data and code supplements) and task 5.5 (main results in an annotated CSV). Task 5.3 will generate computational reproducibility reports published as text (PDF). After applying the series of automated screening tools on the manuscripts, task 5.2 will produce screening reports as PDF, and collect the results of the screening in a spreadsheet (CSV).
- The iRISE consortium will further be gathering, collecting and producing **traditional quantitative (empirical) data** that will be stored in non-proprietary spreadsheets (CSV). More specifically, task 1.4 will use a collection of preexisting data in an annotated CSV table. Any newly collected data from this task will be stored as CSV too. The iRISE SOLES (task 2.1) will be trained on a training dataset which will be an annotated CSV. WP3 will gather empirical data in CSV spreadsheets for task 3.1 (data collected during EDI – equity, diversity and inclusion – support), task 3.2 (data collected during a stakeholder analysis) and task 3.6 (data collected during a bibliometric data analysis). Some intervention studies planned in WP5 will also produce empirical data. Notably, task 5.2 will collect information on papers included in the intervention studies and the study results in a CSV spreadsheet. Task 5.3 will collect data during the computational reproducibility review. Task 5.5 will collect the aggregated results from its simulation study in a CSV spreadsheet. Task 5.7 will collect bibliometric "demographic data" (related to EDI) on papers that are part of its validation testing.
- The iRISE-SOLES (task 2.1) will be running based on a full-fledged structured **SQL database** because such a database is more stable than a CSV table and regular updates of the database are easier.
- iRISE will further collect **contact data** to interact with participants in surveys, interviews, Delphis and focus groups as for tasks 3.2, 4.1, 4.2 and 5.4; intervention studies (tasks 5.2 and 5.3) and its stakeholder forum (task 6.7). Contact information will be collected and stored in spreadsheets (CSV and XLSX).
- Some of the iRISE WPs will conduct surveys, focus groups and interviews, thus collecting **qualitative and quantitative survey data and interview data**. In such, the surveys and/or focus groups planned in WP3 will collect qualitative data with survey results as CSV, as well as potential audio recordings (MP3) and transcripts of the recordings (as text files, stored as PDF). On top of survey results, the Delphi planned in task 4.1 will record the final coordination or consensus meeting using Microsoft Teams (MP4). The transcript of that meeting will be collected as text files (stored as PDF). The focus group meetings in task 4.2 will be recorded as video and audio if the focus group is held online, and only audio if the focus group is held in person. Transcripts will be collected as text files (stored as PDF). The same type of data will be collected during the surveys and interviews during task 5.4. All raw qualitative data will be summarised and aggregated into CSV spreadsheets and/or lists in text format (stored as PDF).
- iRISE will produce further **texts as output**, in PDF format. These outputs include the glossary

with definitions of reproducibility (task 1.1), a framework to evaluate interventions to improve reproducibility (task 2.3), a dissemination plan (task 6.1), a data management plan (task 6.2), an implementation guide (task 6.3), a policy briefing (task 6.4), train-the-trainer course material (task 6.6), and diverse summary reports and lists, output syntheses and other frameworks (tasks 2.3, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 4.1, 4.2, 5.6).

The data types which iRISE will re-use and which are listed above include spreadsheets (CSVs), R scripts with code, R-packages and tools, and text as PDF.

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The data generated and re-used directly relate to the iRISE objectives, and all protocols will transparently outline our work processes. Literature reviews will help the consortium understand the depth of literature on reproducibility definitions and metrics, and intervention studies to improve reproducibility. The data generated in literature reviews will further efficiently train and be used by iRISE-SOLES. Quantitative data will be re-used and gathered to answer the research questions outlined in the iRISE proposal. Qualitative data collected via Delphi method surveys, surveys, interviews and focus groups will complement existing data to gain insights on prioritisation of interventions, facilitators and barriers of implementation, and underrepresented groups. Code and other computational tools will allow us to extract the most useful information from different data sources and represent our results adequately.

What is the expected size of the data that you intend to generate or re-use?

Most data generated and re-used will be small. We expect individual data outputs (text, spreadsheets, code scripts) to stay well below 100MB in size. The SQL database for the iRISE-SOLES is expected to be at most 1GB large. The totality of the raw qualitative data might exceed this size, as it includes video and audio recordings. However, also due to sensitivity considerations, the number of recordings will be small, and the size of the data will not hinder them from being stored on the iRISE SharePoint.

What is the origin/provenance of the data, either generated or re-used?

In general, the origin/provenance of each individual data output will be outlined in detail as part of our minimal metadata standards. This includes concrete workflows put in place to collect the data. Such information will also be part of the preregistered protocols. Specifically, survey data will be collected via SurveyMonkey and/or Qualtrics. Qualitative data will be aggregated in NVivo (standard software), and the bibliographic data will be annotated in open-source literature review software (SyRF used in tasks 1.1, 2.2). Two of the intervention studies (task 5.2 and 5.3) will rely on a collaboration with publishers which will provide access to submitted manuscripts.

To whom might your data be useful ('data utility'), outside your project?

All data collected during the iRISE project will be useful for any research group that intends to reproduce or build on our findings. This is particularly true for the protocols outlining our work processes and methods, the code and tools we build and make openly available, but also the data

collected and aggregated in spreadsheets. More specifically, the glossary with reproducibility definitions and metrics will condense information and therefore be useful for the entire research community. The data collection of reproducibility projects and other examples will be a great resource for any other (follow-up) project in need of empirical data on reproducibility. The data collected in WP3 will be useful to inform future research projects and consortia on how to integrate equity, diversity and inclusion (EDI) dimensions. Specific iRISE data outputs are meant to help and train stakeholders to implement interventions to improve reproducibility. To further facilitate dissemination of the iRISE (data) outputs and results, some outputs will be integrated into the Embassy of Good Science.

FAIR data

2.1. Making data findable, including provisions for metadata: Will data be identified by a persistent identifier?

Yes, all data outputs which will be openly available will have a persistent digital object identifier (DOI). The DOI will be provided through a repository of choice (the Open Science Framework or Zenodo). Code will be version controlled via Git and/or linked to OSF or Zenodo to be allocated a DOI. The data that must remain closed (mentioned below) will not have a DOI.

2.1. Making data findable, including provisions for metadata: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Yes, all data outputs gathered by iRISE will be accompanied by rich, comprehensive metadata. Even if the data itself is closed-access, the data will have publicly available metadata (most likely on OSF or Zenodo). This excludes contact information of participants and stakeholders. To ensure that all iRISE partners systematically provide the right type of metadata, the iRISE project management handbook includes a description of the required minimal metadata standards. For this minimal standard, the following elements (with description) are required:

- **Title**: title describing the data output at hand.
- **Principal Investigator or Creator**: the main person(s) responsible for the intellectual content, with affiliation(s).
- **Contributor(s)**: any other person(s) who contributed to the data output with affiliation(s).
- **Funding**: funding source of the project leading to the data output (iRISE and additional funding sources must be acknowledged here).
- **References and citations**: Citations to relevant work or other objects/material leading to the data output or using the data output. Only cite those articles or material that are important for the data output to be reusable and interpretable. Specifically, if applicable, cite any software or material needed to interact with the data.
- **Summary | Description**: A textual description of the aims of data collection and a summary of the data output itself (in the form of a short abstract).
- **Keywords**: List of relevant keywords making the metadata findable.

- **Coverage**: when and where was the data collection - or the project - started and when was it finalized.
- **Date of publication**: Date of data deposition (first – and new versions)
- **Unit of observation**
- **Population**: information on the population of interest represented or targeted in the data output.
- **Data type and format**: information on the type and format of the data collected.
- **Sampling and weighting**: information on whether any sampling or weighting was used in the data acquisition, and if so, which type or method of sampling and/or weighting was used.
- **Mode of Collection**: information on how the data was collected, on the method used for data collection.
- **DOI**
- **Licenses and restrictions**
- **Ethical considerations**: if ethical approval was needed and acquired, the metadata should link or cite the ethics approval.
- **Description of variables**: if possible, this should be done in a separate code book or data dictionary.

2.1. Making data findable, including provisions for metadata: Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Keywords will be part of the metadata provided with all the data outputs created by iRISE. Keywords on OSF are referred to as tags.

2.1. Making data findable, including provisions for metadata: Will metadata be offered in such a way that it can be harvested and indexed?

Whenever possible and the repository allows it, the metadata will be supplied in the repository user interface (possible for OSF and Zenodo) which ensures that the metadata is machine-readable.

2.2. Making data accessible - Repository: Will the data be deposited in a trusted repository?

Yes, all data will be made accessible in trusted public data repositories. At the start of the project, iRISE created its open knowledge base on the Open Science Framework where most of the data will be uploaded (specifically, text data including protocols, reports and summaries, smaller spreadsheets, references to code on Git). Larger, more complex data will instead be deposited on Zenodo as it is more suited for complex data. Also working snapshots of Git-repositories will be deposited on Zenodo or OSF to ensure long term preservation and get a DOI. All the iRISE data output, and work in progress

will additionally be archived, versioned and secured on the iRISE SharePoint hosted by the iRISE coordinating partner (Charité). The sensitive data, mainly contact information for participants in various tasks and the raw qualitative data collected in surveys, interviews and alike, will be archived on the iRISE SharePoint and password protected. Only task leads and key individuals will have access to the password. An exception will be the contact information and data needed for the task 5.3 intervention study based on computational reproducibility review for which the data will stay at Karolinska Institutet, and the data collected in task 5.2 which will be stored on a separate Charité server operated by the task-leading partner. This will simplify the data transfer agreement with the publisher.

2.2. Making data accessible - Repository: Have you explored appropriate arrangements with the identified repository where your data will be deposited?

The Open Science Framework is a free and reliable service. For example, the OSF FAQs explain that "the OSF database is backed up via streaming replication 24 hours a day, and incremental restore points are made twice daily". Already at the time of proposal submission, the iRISE partners created an OSF project which will now act as the iRISE open knowledge base. Additionally, larger and more complex data will be uploaded on Zenodo, as it is a more appropriate repository to deal with complex data. Working snapshots of the iRISE Git-repositories will be uploaded to Zenodo to ensure long term storage of the code and software produced. Zenodo is another free service with easy access to all partners and all interested in the project and its outputs. Zenodo also has a comprehensive strategy for longevity of the data uploaded. We will create a Zenodo community to ensure all generated data is connected to iRISE and easily findable and identifiable as such. Finally, another free and accessible repository or service used is Git. To increase findability, all Git-repositories will be linked to the iRISE open knowledge base on OSF. Linking to OSF and/or uploading a snapshot to Zenodo also allows the code and software to have a DOI.

2.2. Making data accessible - Repository: Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Yes, OSF and Zenodo allow the data to be assigned a persistent digital object identifier (DOI). Git does not provide this service. However, all iRISE Git-repositories will be linked to from the iRISE Open Knowledge Base on OSF or Zenodo, which allows the code and software generated to have a DOI. Working snapshots of the iRISE Git-repositories will also be assigned a DOI via Zenodo.

2.2. Making data accessible - Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

Apart from sensitive data, all data generated by the project will be made openly available. The sensitive data includes contact information for the stakeholder forum, the Delphi panel (task 4.1), the focus groups (task 4.2), the intervention study based on computational reproducibility review in task 5.3 and the intervention study based on the EQIPD quality system in task 5.4. Further, the raw survey results (gathered in tasks 4.1 and 5.4) will be closed and only the aggregated anonymized version of the data is shared openly. The sound and/or video recordings of the final consensus meeting of the

Delphi (task 4.1) and the focus group meetings (task 4.2) will not be made openly available. The same is true for any transcripts of the discussions during these meetings or interviews (from task 5.4). The raw SQL database for the SOLES will be restricted access, to ensure its implementation is not interrupted. However, (monthly) read-only snapshots of the database will be openly available and can be reused by other projects. Additionally, WP1 will preregister predictions on the efficacy of the interventions tested in WP5 (task 1.5), which will be closed until the completion of the respective intervention studies. The study protocol for task 5.2 will be closed until the study is concluded to avoid the participants being biased.

2.2. Making data accessible - Data:

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

No embargo will be applied to our output data. Either the data is open and accessible right after it is generated and uploaded to a repository at completion or submission of the respective task, or it is closed and protected, due to its sensitivity (only true for contact information and raw qualitative data). Exception to this rule are the predictions of the efficacy of the intervention studies (planned in task 1.5), which will be closed until completion of the respective intervention studies, and the study protocol for task 5.2, which remains closed until task 5.2 is concluded.

2.2. Making data accessible - Data:

Will the data be accessible through a free and standardized access protocol?

Yes, all data deposited on OSF, Zenodo, and Git will be free and directly accessible to everyone.

2.2. Making data accessible - Data:

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

The contact information will be password protected, archived, and encrypted on the iRISE SharePoint hosted at Charité and at Karolinska Institutet during and after the completion of the project. Encryption keys will be deposited on the iRISE SharePoint (also password protected). Contact data will only be accessible to those researchers directly working on the respective task who have additionally signed a data sharing or processing agreement. This is also true for the raw qualitative data from surveys, interviews, and alike. The iRISE Steering Committee will have full access to the passwords and share only specific passwords with the researchers directly involved in the respective tasks. The anonymized and aggregated version of the data will be directly accessible to all iRISE partners (via the iRISE SharePoint).

2.2. Making data accessible - Data:

How will the identity of the person accessing the data be ascertained?

Any person will be able and allowed to access our open data via OSF, Zenodo and Git. To access the data on the iRISE SharePoint and the data hosted by Karolinska Institutet for task 5.3 or the separate

Charité SharePoint for Task 5.2, a personalized login will be needed. All iRISE partners will have access to the iRISE SharePoint. Where applicable, for sensitive data, including contact information and raw qualitative data, access will be outlined in the task-relevant data transfer agreements or study protocols.

2.2. Making data accessible - Data:

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

There is no need for a data access committee. It is clear from the project proposal which partners are responsible for the underlying tasks and require access to the data. Additionally, data sharing and/or processing agreements will be signed clarifying any open concerns, specifically for the tasks collecting sensitive and contact data and the intervention studies who receive data from a publisher.

2.2. Making data accessible - Metadata:

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

Yes, metadata will be openly available and licensed under a public domain dedication CC0 (from OSF and Zenodo). As described above, the metadata will be comprehensive and ensure future users can access and re-use the data.

2.2. Making data accessible - Metadata:

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

We do not plan to close access to or delete any iRISE generated data. Hence, the data and metadata will be available and findable as long as OSF, Zenodo and Git continue hosting it, free of charge. In their FAQs, OSF explains that they have a preservation fund ensuring sustainability of the hosted data for 50+ years in the case of closure of the OSF offices. Zenodo writes that “In the highly unlikely event that Zenodo will have to close operations, we guarantee that we will migrate all content to other suitable repositories.” Finally, Git is a version control system which will undoubtedly not be discontinued any time soon. To be on the safe side regarding Git, which does not have a clear strategy for long term preservation, snapshots of working versions of the iRISE Git-repositories will be deposited on Zenodo.

2.2. Making data accessible - Metadata:

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

Code and software are – according to our definition (see above) – part of data. All data will be deposited and archived in a non-propriety format, including CSV for spreadsheets, R, Python and CSS scripts for code and software, PDF, XML and HTML for text documents, MP3 and MP4 for audio and video recordings. All software needed to interact with the data will be open-source and referenced or cited in the metadata. If specific figures or illustrations need to be appealing to a broader audience,

proprietary software might be used which will however not prevent the underlying data from being accessible with open-source software. WP4 and WP5 plan to use established software to collect and interact with sensitive qualitative data (specifically NVivo, SurveyMonkey and/or Qualtrics).

2.3. Making data interoperable:

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

Protocols and subsequent publications will follow the appropriate reporting guidelines. Whenever possible, the glossary and reports will use controlled vocabulary, while, for our area of research there is no “community accepted” vocabulary. All iRISE partners will follow the metadata standard outlined above (in 2.1. Making data findable, including provisions for metadata) to ensure that it is comprehensive and includes all relevant information. Where applicable, data dictionaries and code books will complement the metadata. Analytical code and software will be well documented and follow well-known style guides. READMEs and explanatory examples will further help the interoperability of code and software.

2.3. Making data interoperable:

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

Documentation of the generated data will use common, standardized and controlled vocabulary. Protocols and manuscripts will follow the appropriate reporting guidelines (including PRISMA for literature reviews, ADEMP for simulation studies, STROBE for observational studies, COREQ for mixed methods). The open iRISE glossary will further outline the definitions of the types of reproducibility which will be used throughout the project and might contribute to the development of a community accepted vocabulary.

2.3. Making data interoperable:

Will your data include qualified references[1] to other data (e.g. other data from your project, or datasets from previous research)?

[1] A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)

Yes, we plan to ensure all our data is linked or referred to the other iRISE outputs. All data not directly uploaded to the iRISE knowledge base will be linked from the open knowledge base to ensure interested people have one main source of information to consult.

2.4. Increase data re-use:

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

All studies and analyses planned in iRISE will have a study protocol which will be preregistered on OSF and be part of the iRISE knowledgebase. The data generated from each of these studies and analysis will be consecutively added to the same OSF component. As described above, data, code and software will be complemented with readmes, code books, data dictionaries, examples, workflows etc. to ensure interoperability and re-use.

2.4. Increase data re-use:

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

All data generated by iRISE which is not sensitive (contact information and raw qualitative data) or subject to a specific data transfer agreement (with a publisher) will be openly available in the public domain and licensed under CC-BY 4.0 International. Only the predictions of the efficacy of the intervention studies produced and preregistered in task 1.5 and the study protocol for task 5.2 will be closed until the respective intervention studies are completed (they will however have open metadata and a license). The iRISE-SOLES SQL database will be closed access to guarantee the SOLES is executing smoothly, but (weekly) snapshots of the database will be available from OSF.

2.4. Increase data re-use:

Will the data produced in the project be useable by third parties, in particular after the end of the project?

Yes, all open data will remain on the selected repository and any third party will have access, also after completion of the project (for as long as the repositories operate).

2.4. Increase data re-use:

Will the provenance of the data be thoroughly documented using the appropriate standards?

Yes, part of the metadata will be a documentation of the provenance of the data. Additionally, the preregistered protocols will outline the data generating process.

2.4. Increase data re-use:

Describe all relevant data quality assurance processes.

To ensure the quality of the methodology and processes used, the protocols will be written by the task leads but reviewed by the WP responsible for the task and, if judged relevant, by the whole consortium. Where applicable, stakeholders and/or other consortia or groups from outside iRISE will be

consulted at an early stage. For the analytical code and software, code review will be arranged and facilitated through the iRISE consortium. For data spreadsheets systematic data quality checks will be put in place. Making the data openly available as early as possible will further facilitate data review by other members of the research community. Specifically, for the iRISE-SOLES specialized quality assurance processes will be established: unit-tests for each function as part of the automation tools, the training dataset will be annotated by at least two reviewers, the SOLES SQL database will undergo regular tests and back-ups, and the RShiny app will include regular testing. iRISE will, whenever possible, follow standard, well-established methods, guidelines and standards.

2.4. Increase data re-use:

Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

As explained above, we consider data to include all information or material needed to reproduce the iRISE results and outputs. This includes datasets, software, code, protocols, workflows, tables, images, videos, transcripts, articles. Links between the different data outputs will become apparent from metadata and documentation.

Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

All questions answered above already refer to other research outputs since they are part of our iRISE data output definition.

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

This was already answered above.

Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?

Most of the data can be made openly available and as FAIR as possible without any cost related to storage, archiving, etc. Much of the data will be backed-up on the iRISE SharePoint hosted by Charité, which will also be without cost for the entirety of the project and beyond. For some specific tasks, other institutional servers will be used free of charge to facilitate data transfer agreements. To keep the iRISE-SOLES running, a team member needs to regularly check that the scripts and cloud computing infrastructure are running. The workload here is estimated to be about two hours per month. Additionally, the SQL database will be running on Amazon web services, which is subject to a fee. Some of the functionalities of the iRISE-SOLES are based on Large Language Models (LLMs), which also comes with some costs.

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

The ongoing cost of maintaining the iRISE-SOLES database and web application after the completion of iRISE is minimal in the broader context (approximately £50 per month). The responsible partner, the University of Edinburgh, already hosts several other SOLES project databases on Amazon Web Services and associated web applications on shinyapps.io. We therefore consider it highly likely that this partner can provide long-term support for the platform.

Given the rapid pace of development in this field, the cost and choice of LLMs for iRISE-SOLES will be reassessed in light of the evolving technological landscape prior to project completion. At that stage, a decision will be made - together with the responsible partner’s institution - on whether to transition to a lower-cost open-source model, continue using the ELM service, or adopt an alternative proprietary solution, potentially supported by additional external grant funding.

Who will be responsible for data management in your project?

The task lead will be responsible for the management of the data generated by their task. The WP leads are responsible for keeping an eye on data management. The authors of this DMP (R. Heyard and G. Currie) oversee the iRISE data management and can be consulted for help if necessary.

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

The iRISE consortium plans to forge partnerships with stakeholders who use the iRISE-SOLES platform to support this effort in the long term. All underlying code for the user interface will be openly available to ensure a copy can be generated at any time if an issue were to affect the continuity of the platform. For all other data outputs, long-term preservation will be facilitated through the respective data repositories (OSF, Zenodo). Specifically, since Git does not have a strategy for long-term preservation, all Git-repositories will have a working snapshot uploaded to Zenodo.

Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

All data outputs will be backed up on the iRISE SharePoint, to which only registered iRISE partners have access. The sensitive data will be deposited on the iRISE SharePoint and will be encrypted, with the encryption key also on the SharePoint. The access to this data, specifically to the encryption key, will be strictly controlled (e.g., only partners involved in the specific tasks will have access). For the tasks collaborating with publishers, data transfer agreements will outline the storage and access in detail. The iRISE-SOLES SQL database will be backed-up regularly, and snapshots will be uploaded to OSF and made openly available. The iRISE-SOLES' will use LLMs through the Edinburgh (access to) Language Models (ELM), which provides safer, more secure access to OpenAI's GPT-4 (and other models), in line with the University’s AI ethics principles. ELM operates under a Zero Data Retention agreement with OpenAI, ensuring that no data submitted through the platform is stored or reused, and all interactions remain confidential.

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

‘Section 4. Ethics self-assessment of the iRISE Description of the Action’ discusses ethical dimensions and the compliance with ethical principles in detail. Some tasks, analyses and studies planned in iRISE will need ethical approval. This includes the qualitative and mixed-methods research planned by WP4. The strategy for ethics approval will be discussed in the respective protocols. The same is true for some of the intervention studies planned in WP5 (task 5.2, task 5.3 and 5.4). Overall, any ethics and/or legal considerations required were discussed above, e.g., in relation to sensitive data.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

Whenever personal data is collected in questionnaires or alike, participants will be informed on what types of data is collected and how and for how long the data will be preserved. For example, informed consent documentation will be created for the Delphi consultation process and the focus group discussion in WP4. Further considerations will be outlined in the ethics sections of the protocols for the studies collecting personal data.

Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

N/A