⟨ EQUAL|TY ⟩

Efficient QUantum ALgorithms for IndusTrY

WP7 Management

D7.3 Data-management plan

Version: 1.0

Date: 28/03/2023

AIRBUS   *Capgemini*   DA VINCI LABS

Fraunhofer ENAS   DLR Deutsches Zentrum für Luft- und Raumfahrt German Aerospace Center   Inria

Universiteit Leiden The Netherlands   PASQAL

Document control

Project title	Efficient QUantum ALgorithms for IndusTrY
Project acronym	EQUALITY
Call identifier	HORIZON-CL4-2021-DIGITAL-EMERGING-02
Grant agreement	101080142
Starting date	01/11/2022
Duration	36 months
Project URL	http://equality-quantum.eu
Work Package	WP7 Management
Deliverable	D7.3 Data-management plan
Contractual Delivery Date	M6
Actual Delivery Date	M5
Nature[1]	DMP
Dissemination level[2]	PU
Lead Beneficiary	DLR
Editor(s)	Linus Scholz, Michael Epping
Contributor(s)	
Reviewer(s)	Wael Yahyaoui
Document description	This deliverable will describe the strategies and measures of data management; how the data will be managed, analysed, described, and stored, what standards will be used, how the metadata will be handled, and how the data will be protected.

---

[1]R: Document, report (excluding the periodic and final reports); DEM: Demonstrator, pilot, prototype, plan designs; DEC: Websites, patents filing, press & media actions, videos, etc.; DATA: Data sets, microdata, etc.; DMP: Data management plan; ETHICS: Deliverables related to ethics issues.; SECURITY: Deliverables related to security issues; OTHER: Software, technical diagram, algorithms, models, etc.

[2]PU – Public, fully open, e.g., web (Deliverables flagged as public will be automatically published in CORDIS project's page); SEN – Sensitive, limited under the conditions of the Grant Agreement; Classified R-UE/EU-R – EU RESTRICTED under the Commission Decision No2015/444; Classified C-UE/EU-C – EU CONFIDENTIAL under the Commission Decision No2015/444; Classified S-UE/EU-S – EU SECRET under the Commission Decision No2015/444

Version control

Version	Editor(s) Contributor(s) Reviewer(s)	Date	Description
0.4	Linus Scholz	10/02/2023	Intermediate document finished.
0.5	Wael Yahyaoui	13/02/2023	Intermediate document approved by the reviewer.
0.8	Linus Scholz	20/03/2023	Document finished by editor.
0.98	Wael Yahyaoui	27/03/2023	Document approved by reviewer.
1.0	Wael Yahyaoui	28/03/2023	Document released by Project Coordinator.

Abstract

A quantum revolution is unfolding, and European scientists are in the lead. It is time to take decisive action and transform our scientific potential into a competitive advantage. Achieving this goal will be critical to ensuring Europe's technological sovereignty in the coming decades. EQUALITY brings together scientists, innovators, and prominent industrial players with the mission of developing cutting-edge quantum algorithms to solve strategic industrial problems. The consortium will develop a set of algorithmic primitives applicable to various industry-specific workflows. These primitives include differential-equation solvers, material-simulation algorithms, quantum optimisers, and quantum machine learning. To focus our efforts, we target eight paradigmatic industrial problems. These problems will likely yield early quantum advantage and pertain to the aerospace and energy-storage industries. They include aerodynamics simulation, battery- and fuel-cell design and optimisation, battery-material discovery, space-mission optimisation, and space-data analysis.

We aim to develop quantum algorithms for industrial problems using real quantum hardware. This requires grappling with the limitations of present-day quantum hardware. Thus, we will devote much of our efforts to developing strategies for optimal hardware exploitation. These low-level implementations will account for the effects of noise and topology and will optimise algorithms to run on limited hardware. EQUALITY will build synergies with Quantum-Flagship projects and Europe's thriving ecosystem of quantum start-ups. Use cases will be tested on quantum hardware from three of Europe's leading vendors and two HPC centres. The applications targeted have the potential to create billions of euros for end-users and technology providers over the coming decades. With EQUALITY, we aim to play a role in unlocking this value and placing Europe at the centre of this development. The project gathers nine partners and has a budget of €6M over three years.

Disclaimer

This document does not represent the opinion of the European Union or European Commission, and neither the European Union nor the granting authority can be held responsible for any use that might be made of its content.

This document may contain material which is the copyright of certain EQUALITY consortium parties and may not be reproduced or copied without permission. All EQUALITY consortium parties have agreed to the complete publication of this document. The commercial use of any information in this document may require a license from the proprietor.

Neither the EQUALITY consortium nor a certain party of the EQUALITY consortium warrants that the information contained in this document are capable of use nor that the use of the data is free from risk. It accepts no liability for loss or damage anyone suffers using this information.

Acknowledgement

Table of contents

List of abbreviations

D	Deliverable
GDPR	General Data Protection Regulation
HPC	High-Performance Computing
HTTPS	Hyper Transfer Protocol Secure, an encrypted form of web-based communication that is now standard across websites
PDF	Portable Document Format, a popular format used for documentation such as reports and whitepapers
SVN	Apache Subversion
WP	Work Package

List of tables

1. EXECUTIVE SUMMARY

This document is a deliverable of the EQUALITY project, funded under grant agreement number 101080142.

This deliverable, D7.3, is part of work package 7 (WP7) and details the data-management plan for the EQUALITY project. Throughout the project's lifetime, a variety of data sources will be used and generated, and as such, the correct handling, protection and archiving of these are essential. This document employs a standard data-management template to describe the significant data sets envisaged for the EQUALITY project to be used throughout its lifetime. For each data set, a detailed description of the project plan around handling, sharing, archiving, and backup will be explored. The document concludes with miscellaneous issues around data management, including those driven by specific legal frameworks. The data management plan is a live document; its version and date should be tracked. As such, this document is version 1.0, April 2023, of the data-management plan.

2. INTRODUCTION

The data-management plan identifies the critical forms of data that the project team will create and manage. For each of these data forms, guidance is provided on how the data should be formatted, marked up and curated, aiming to ensure that:

1. The data should be findable, accessible, interoperable, and reusable (FAIR) by others within and outside the project (as appropriate).
2. It is managed securely based on the value and sensitivity of the data.
3. Data is protected from accidental loss.

Further, the expected use of each data form is documented, along with guidelines as to applications, media, and tools that are well-suited to the anticipated usage. All members of the EQUALITY project team should be familiar with this (and subsequent versions of the) data-management plan and adhere to the guidelines in their project-related work. The work-package WP7 (management) and the task T7.4 coordinator maintain and keep this document current.

3. DATA SUMMARY

3.1. Data collection – shared project materials

3.1.1. Dataset description

During the project, various documents will be prepared for internal use, including, *e.g.,* meeting minutes, briefing notes, team-member contact information, and process documents. This material will be held in the project's document repository (Microsoft SharePoint), which implements fine-grain access and version control. This is accessible via a web interface using HTTPS. The repository is a tree-like hierarchy of folders organised by work packages (WPs 1 to 7) and governance entity. Authentication is done via Microsoft Active Directory and is encrypted.

3.1.2.   Standards and metadata

The standard format of project-reporting documents, including deliverables, are those provided by Microsoft Office (*i.e.*, docx, pptx, *etc.*). Publishing documents for public consumption will be done in PDF, and writers of technical collateral, such as conference and journal papers, can use any technology (*e.g.*, Word, Latex) conforming to the publisher's requirements: appropriate metadata track authorship and contributions, last modified time, and versioning.

3.1.3.   Data sharing

All project materials will be treated as internal and protected by access-control mechanisms by default. The PDF format should be used to share materials more widely (*e.g.*, final versions of slide presentations, white papers, and dissemination material). Any public sharing of project materials must align with the Consortium Agreement's terms.

Project material containing personally identifiable information should only be distributed within the project team if there is a genuine business interest (*e.g.*, noting the authors of a document and their institutional affiliation). Project material containing personally identifiable information should typically be shared within the EQUALITY project as a link to the copy held in Microsoft SharePoint or via another file-sharing service rather than as an e-mail attachment.

3.1.4.   Archiving and preservation (including storage and backup)

Collective project materials in documents, plans, notebooks, messages, and research data are maintained in a Microsoft Teams environment under the Office365 tenancy of the Fraunhofer Gesellschaft. All data processing in the Office365 cloud is subject to a data-processing contract between the Fraunhofer Gesellschaft and Microsoft. Local working directories are located on the storage systems of the respective institutes, where local IT measures to safeguard against data loss and third-party access are in place. Backups of the collectively used Microsoft Teams-accessible project data and local working directories are created regularly onto available storage at Fraunhofer ENAS, which is limited in visibility to the project team, to safeguard against data loss and service degradation. Access to these project materials is secured using two-factor authentication and access-management systems. Selected project materials can be published using central data repositories, like Zenodo or Fraunhofer-Publica.

3.2.   Data collection – research data

3.2.1.   Dataset description

Measuring the performance of software, algorithms, and new methods developed by the project is critical in carrying out the work and ensuring the project meets its goals. Benchmark measurements (often run-time, scalability, and memory efficiency) will be recorded for each significant revision to project software. These will be curated in an agreed form for easy comparison and cross-reference. In addition, any programs or scripts (*e.g.*, used for the analysis of these raw data files and needed for reproducibility) will also be stored.

Native output files, raw data files, processed data files, and the source code for analysis programs will most likely be stored in each partner's source-code repository (*e.g.*, Git, SVN) but, in some circumstances, might be uploaded to the EQUALITY Teams folder.

As much as possible, operations will be automated, using scripts to, *e.g.*, record the precise details of the system environment, build the required software, execute the chosen benchmark programs, extract the resulting measurements, and perform analysis using these results. An operator will manually trigger benchmarking when a significant revision to project software has been created. These scripts will be treated as any other project software.

⟨EQUAL|TY⟩

3.2.2. Standards and metadata

At the time of each significant software run, the operator will use a script to take careful note of the experiment parameters (including context information such as names of those involved, date and time, plus specific information about the versions/ source/ build parameters of any software used in the experiment, the hardware/ platforms used, and any options or switches passed to the benchmark process). This metadata will be recorded in a suitable text file and added to the respective source-code repository simultaneously as a significant revision to the project software to maintain the association. It should be noted that, for reproducibility purposes, a version should be associated with the data sets used for a specific run so that this can easily be tracked in the future, even if the data themselves have changed.

3.2.3. Data sharing

We envisage that selected results will be disseminated for public consumption in terms of both project deliverables and research papers. As part of this, likely, the performance characteristics of the individual applications that make up the workflow of a specific urgent event response will also be publicly accessible. The project team may choose to share tools with other groups outside of the EQUALITY project team, *e.g.*, to enable the checking of project results or comparative investigations to be undertaken, according to the terms of the Consortium Agreement. The EQUALITY website can be used for the public sharing of data. Table 1 summarises the primary research data generated and/or collected through the EQUALITY project.

3.2.4. Archiving and preservation (including storage and backup)

Source code, artifacts (binaries, libraries, containers), and benchmarking results will be stored in the respective partners' source-code repositories, which are most likely versioned so evolutions can be captured over time. For important code bases and artifact repositories, backups are created using repository mirroring.

Table 1: Overview of generated and/or collected research data through the lifetime of the EQUALITY project.

Research data that are generated and/or collected during the project	Accessibility	How data will be shared/exploited	Formats
Scientific documents	Public	Public reports, research publications in open-access journals, patent applications	Written documents (.pdf, .doc, .docx)
Experimental data	Public/confidential, shared with partners	Corresponding experimental data of public deliverables/reports or research publications will be published in open access	Experimental data in raw format (.csv, .xlsx), experimental report (.docx, .pdf, .txt), experimental images (.tiff, .png, .bmp, .jpg, .eps, .mp4)
Source code and artifacts	Public/confidential, shared with partners	(Open-access) source-code repositories (*e.g.*, Git, SVN)	Code (.cpp, .py), binaries (.bin, .exe)

3.3. Data collection – e-mail list

3.3.1. Dataset description

The project maintains a mailing list of e-mail addresses (and, in some cases, names) of internal people. This mailing list is held on EQUALITY's e-mail-list service, which is securely managed by Da Vinci Labs and cannot be updated externally to this. Whilst this mailing list is

publicly reported, project members can't view any details about subscribers. The mailing list is administered by and under the responsibility of Da Vinci Labs.

3.3.2. Data sharing

The subscriber list contains personally identifiable information (e-mail addresses and potential names) and is subject to GDPR. A privacy statement should be accessible and included (or linked) in each message posted to the list.

Project-team members may only make posts to the mailing list. They should only contain important personal information with a legitimate business interest (*e.g.*, presenters' names, the author lists for papers). Capgemini must explicitly approve all posts to this list.

3.3.3. Archiving and preservation (including storage and backup)

The archiving of messages to the e-mail lists and the backing-up of subscribers must be considered. Personal data processing takes place to deliver updates and news regarding the project. A list of e-mail addresses is maintained for the project and one year beyond.

4. FAIR DATA

4.1. Making data findable, including provisions for metadata

All open EQUALITY documents and data will be made findable with persistent links and a clear metadata schema.

We will use the ORCID reference (Open Researcher and Contributor ID) for researchers in our team. ORCIDs are valuable because they connect all a researcher's output regardless of changes to the author's name or affiliation. Having all the scholar's output linked together makes their output more easily quantifiable. We will use the Direct Object Identifier (DOI) for digital research materials. A document's DOI will remain fixed even if the location or metadata of the document changes and can be assigned to articles, datasets, or any other shared project output.

For public project documents, the naming convention will be based on the hierarchical structure of the project. The parent file will be the work package's title, and each work-package file will contain a minute file and files for each task within the work package. The minutes file will contain the meeting minutes for the work package. In contrast, the task files will have documents about that task's activities and may include an additional minute file containing meeting minutes specific to that task.

4.2. Making data openly accessible

As many EQUALITY results as possible will be made openly accessible according to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020 [1]. These will be made available through an open-access repository, and those produced as direct outcomes of the project will be published as detailed in section 3.2.4 above.

Open access to peer-reviewed scientific publications will be performed by publishing either in green or gold open-access journals and announced on the EQUALITY website and the OpenAIRE portal[3]. The process can be automated by depositing the publication in an

---

[3] https://openaire.eu/

⟨EQUAL|TY⟩

OpenAIRE-compliant repository by the authors in the case of green open access or by the scientific publisher in the case of gold open access.

If the repository is not fully OpenAIRE-compliant, the publication or data must be linked at https://www.openaire.eu/participate/claim with the European Commission as a funding agency.

Green open access without publication fees is preferred for disseminating EQUALITY results, *e.g.*, using the Open Research Europe services[4] the EC offers as a high-quality publication venue for H2020 and Horizon-Europe funded projects. Journal publications must guarantee permission to self-archive the published article at most six months after publication. However, due to the journal's visibility and acceptance, it may be considered to publish in a gold open access, made available as open access by the publisher. In this case, publication fees are eligible and can be reimbursed as project costs by the EQUALITY partners.

The project will use GitLab to store, publish, and collaborate on open, machine-readable datasets. GitLab will also be used as the primary data repository. Zenodo will be used for project data identifiers. Frictionless will describe, extract, validate, and transform tabular data.

4.3. Making data interoperable

As for the re-use and interoperability of the data, we will use several types of metadata as required: descriptive metadata to help users with the discovery and identification of a resource (*e.g.,* Title, Abstract, Author, and Keywords), administrative metadata to manage an information resource (*e.g.,* provenance, processing, and rights information), structural metadata to describe the relationship between parts of an information resource, especially digital resources which often consist of multiple files, technical metadata to document technical attributes of digital objects or resources, and preservation metadata to support and demonstrate the long-term digital preservation of an information resource or digital object.

4.4. Increase data re-use (through clarifying licences)

All our documentation will include basic information about the data that allow for its correct interpretation and re-use by ourselves in the future and other researchers (README file, README tab, Data Dictionary, Codebook, Commented Code, Lab Notebooks, as applicable).

Each dataset used in the project will be accompanied by a data license that details the permissions associated with using that dataset. The licenses for each dataset will be stored in their respective repository, where they will be found in the separate dataset's README and metadata file.

5. MISCELLANEOUS CONSIDERATIONS

5.1. GDPR compliance

The project (and its data-management tools such as GitLab) will be fully compliant with the GDPR laid out in Regulation (EU) 2016/679 of the European Parliament and the Council of 27/04/2016 on the protection of natural persons concerning the processing of personal data and the free movement data, and repealing Directive 95/46/EC (20) and respects regulations

---

[4] https://open-research-europe.ec.europa.eu/

on intellectual property rights (IPR) (24). Database access will be entirely anonymous. Users must register an account if they wish to use our open repositories. The user must provide their e-mail address and a password and grant their consent that their e-mail address is processed and used for account-related communications and management. The user is free to revoke their consent at any time, which would lead to the deletion of their account.

5.2.   Impact of the GDPR

In addition to the items considered in Section 2, the project team should always be vigilant about the possibility that activities will be subject to GDPR. In particular, the following actions should be raised with Capgemini in advance:

- Creating a new project mailing list will involve collecting and managing personal information (e-mail addresses and possibly names).
- Circulating surveys and web forms: since this may involve the collection, handling, and potential transmission of personally identifiable information, as well as the use of third-party services.
- Collection of website analytics data: Google Analytics, which is used to monitor traffic to the project website, can involve the collection and transmission of personally identifiable information (*e.g.*, specific location information about visitors and IP addresses of web clients). The following must be done concerning Google Analytics:
    - Transmission of personally identifiable data to Google Analytics should be disabled.
    - IP anonymisation should be enabled in Google Analytics.
    - The use of cookies should be avoided, and if using cookies, a Privacy Policy should be provided.

6.   CONCLUSIONS

In this document, we have considered the significant data collections that we will use during the lifetime of the EQUALITY project. For each of these, a detailed description of the data, standards around data format and metadata, the likelihood of sharing data, what limitations are associated with this, and necessary protections, including, *e.g.,* issues around security and archiving, have been explored.

Other miscellaneous concerns have been considered and addressed in section 3. All project members should familiarise themselves fully with this document and ensure that the procedures it contains are followed. Inevitably, as the project progresses, things will change. As such, the WP7 (management) work-package leader and task T7.4 are responsible for keeping this document up to date with the latest information and procedures.

7.   REFERENCES

[1]   European Commission; Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. **2017** https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf. Retrieved 10/03/2023.