Deliverable Report

Author : Thomas Archer / Roberto Troncoso

Date : 29.05.2017

Deliverable	**Deliverable name (Short name)**	**WP No.**	**Lead participant**	**Type**	**Diss. Level**	**Date**
D6.2	Data Management Plan	5	NTNU	Report	Public	M6

TRANSPIRE Data Management plan

Version 1.0 April 25 2017

This plan follows the guidelines presented in the document "Guidelines on FAIR Data Management in Horizon 2020". This document is available to all consortium members and will be updated throughout the course of the project as required.

1.0. Data Summary

- State the purpose of the data collection/generation

  In the course of the TRANSPIRE project data will be generated which will add to our understanding of materials. Data collected will facilitate collaboration in the project and form a resource for future research.

- Explain the relation to the objectives of the project

  The objective of TRANSPIRE is to create a range of new spintronic devices operating up to 2 THz. To create these devices we must consider many possible candidate materials as well as fine tune the defect concentrations, appropriate growth techniques, interface properties, film thickness. During this process we will generate data which is of use to the project but also for further investigation of these and similar materials.

- Specify the types and formats of data generated/collected

  Most data will be in the form of key/value pair and can be easily tabulated. However the degree of investigation will vary from material to material so not all fields in a simple table will be filled. In the key/value pair the values may simply be numbers or block of text, for example explaining an experimental procedure.
  In addition it is expected that some data will be in the form of images, where possible these images will be a marked up vector graphics file with the appropriate meta data included in the image file.

Additional data will be in proprietary formats for example select outputs from a calculation which can be used for further investigation.

- Specify if existing data is being re-used (if any)

  Materials knowledge to date is large we will draw from this data where appropriate. However much of the data is in the scientific literature requiring a significant human effort to retrieve. Where appropriate we will incorporate this data into our database to aid our own investigation and to create a tool for future research.

- Specify the origin of the data

  Scientific literature, data books, ICSG crystallographic database, COD database, Materials Genome projects.

- State the expected size of the data (if known)

  This is difficult to estimate at the beginning of the project, however the storage of images, data sets, and proprietary files can be large. One of the aims of the project is to provide a tool for data aggregation and sharing so we would expect the data set to grow over time. We therefore require that the solution can be easily scaled.

- Outline the data utility: to whom will it be useful

  The data created and aggregated from other sources will be useful to both scientists and engineers requiring knowledge on material properties or the procedures involved in designing materials to meet a specific need.

○ 2.0. FAIR Data

- 2.1. Making data findable, including provisions for metadata
  - ▪ Outline the discoverability of data (metadata provision)

    The primary source will be a linked key/value pair database with all the data appropriately linked, in this case the meta data is by design included in the database.
    Where blobs of data are created these files will be marked up with the information describing the file contents and how it was generated. Where appropriate these files will be linked in the key/value pair database.

  - ▪ Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

    Yes, where possible DOI will be linked in the database of the primary source of data. All publications will be made openly available on open access archive servers, the DOI numbers of these will also be linked in the database.

  - ▪ Outline naming conventions used

    Files will be named on a systematic manor containing the version number.

- Outline the approach towards search keyword

- Outline the approach for clear versioning

  Documents associated with the project, both private and public, will be stored under a version control system (gitlab).

- Specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how.

  Only a hand full of materials data has been standardized, these standards have been created in the crystallogrophy community. We will conform to these standards.

  Keywords are not standardized in the field and there are many terms that refer to the same thing. In the creation of the database we will create an index of keywords and their description.

- 2.2 Making data openly accessible
  - Specify which data will be made openly available? If some data is kept closed provide rationale for doing so

    Data will initially be closed to allow verification of its accuracy within the project. Once verified and published all data will be made openly available. Where possible raw data will be made available however some data requires additional processing and interpretation to make it accessible to a third party, in these cases the raw data will not be made available but we will make the processed results available.

  - Specify how the data will be made available

    A database will be created providing search facilities. This database will link the data together to provide a consistent picture. Where blobs of data are created the database will provide the link and DOI number for the blob.

  - Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

    Documents will be provided via the versioning software git.
    The database will be created under the nosql database software MongoDB. The database will be housed in a cloud insulation. We will enable read access to the backend tools as well as a frontend for simple searches.

  - Specify where the data and associated metadata, documentation and code are deposited.

    The data will be stored in a cloud virtual private server, this will allow the flexibility on hosing databases, git servers, flat file storage and web hosting.

- Specify how access will be provided in case there are any restrictions

  This is a cloud implementation with no limitation on what we chose to provide.

- 2.3. Making data interoperable
  - Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

    We will adhere to the crystal structure standard. We plan on working in collaboration with the "Crystallography Open Database" for linking our data which existing crystal structure data.

  - Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

    Yes a mapping will be defined during the project for interchangeable key words and units of data.

- 2.4. Increase data re-use (through clarifying licenses)
  - Specify how the data will be licensed to permit the widest reuse possible

    After an initial period data will be open access.

  - Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed.

    Data will be made open access within 1 year of its creation, this is to facilitate internal checking and publication.

  - Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why

    Yes the data will be open access.

  - Describe data quality assurance processes

    The metadata will provide all the details of the way in which the data has been generated. Where appropriate links to peer reviewed journal articles will be provided as well as DOI numbers.

  - Specify the length of time for which the data will remain re-usable

    We have provisioned in the project budget for 10 years hosting, the data will be available for this time.

○ 3. Allocation of resources

- Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

We propose to purchase cloud hosting on a virtual private server. A VPS with 150GB storage with "a2 hosting" costs €24.89 per month, for 10 years hosting the total cost is €2986.8. The data will be backed up on servers hosted in TCD. We intend the hosting cost to be paid by the TRANSPIRE project.

- Clearly identify responsibilities for data management in your project

Data management will be done by Thomas Archer.

- Describe costs and potential value of long term preservation of long term data management

Cloud hosting needs to be paid on a monthly basis and the price and our requirement is expected to fluctuate over time. The TRANSPIRE account with appropriate funds should be kept open for until January 2027 to maintain this resource. We estimate that €2986.8 will be sufficient funding. Supplemental funding cannot be guaranteed to maintain this resource but is expected to come from additional projects.

- 4. Data security
    - Address data recovery as well as secure storage and transfer of sensitive data

Data will be synced daily with a server hosted in TCD. The TCD server itself has a zfs raid-z2 file system with daily snapshotting as well as 2 redundant copies of the data.

- 5. Ethical aspects
    - To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

In creating this resource we will not infringe on the copy right held by journals. Any publish image we host must not be a duplicate from a piece of work for which we do not have the rights to publish. All work from this project will be published in open access journals.

- 6. Other