DATA MANAGEMENT PLAN

(Deliverable D1.1)

Project title:

Semiconductor crystal phase engineering: new platforms for future photonics

Project acronym:

SATORI

Grant agreement number:

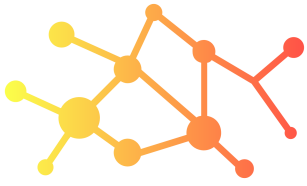
101030927

Author:

Dr. Christopher A. Broderick

SATORI

EU-H2020 MSCA-GF



1. Introduction

1.1. Overview

The MSCA Individual (Global) Fellowship SATORI, henceforth referred to as the “Action”, centres on theoretical analysis of the electronic and optical properties of novel metastable crystalline phases of III-V and group-IV semiconductor materials and their heterostructures. The Action is proceeding via partnership between Tyndall National Institute, University College Cork, Ireland (TNI-UCC, henceforth referred to as the “EU Beneficiary”) and the University of California, Santa Barbara, U.S.A. (UCSB, henceforth referred to as the “Partner Institution”), with cooperation between these institutions in the context of the Action legally underpinned by a Partnership Agreement between these two institutions.

This Data Management Plan is a living document that will evolve throughout the duration of the Action, in response to the types and/or volumes of data generated. The intention of any future revisions to this Plan are to ensure that, where possible, all data associated with the Action are rendered findable, accessible, interoperable and reusable (FAIR), in addition to being secure and persisent. Following the initial version of this Plan (v.1.0.0), revisions are expected to pertain primarily to scientific software developed by the Fellow as part of the programme of research and training associated with the Action, where the choice of input and/or output data and/or file formats (including any associated metadata), is subject to change throughtout the duration of the Action.

1.2. Data summary

The Action is entirely theoretical and computational in nature, involving a combination of (i) applying existing software to perform numerical simulations, (ii) analytical derivation and numerical parametrisation of theoretical models, and (iii) development of bespoke software, with (iii) intended firstly to facilitate input data preparation and/or output data post-processing pertinent to (i), and secondly to provide implementation of numerical simulations based on (ii). All data associated with the Action will be electronic in nature. This Data Management Plan therefore pertains to (i) open-access dissemination of electronic data and/or software associated with technical publications (cf. Sec. 2), and (ii) ensuring the security and persistence of the Fellow’s electronic data and software (cf. Sec. 3).

While significant advances have been made in recent years towards rendering computational materials science data findable, accessible, interoperable and reusable (FAIR) [1], there do not yet exist established overarching metadata standards in the field [2]. It is therefore intended that data and software dissemination associated with the Action will take the form of providing persistent open access to (i) input and output files that are compatible and interoperable with established materials science software, (ii) any related raw pre- and/or post-processed data, and (iii) self-developed software and associated documentation, with either of (i), (ii) and (iii) being released as they relate to individual technical publications. This will be achieved via a combination of secure, persistent archiving (using the Zenodo repository; https://zenodo.org), and open-source version control and distribution software (using git, via the GitHub platform; https://github.com).

Work packages (WPs) WP1 and WP2 of the Action centre on performing first principles calculations based on density functional theory (DFT) and post-processing of associated data, using a combination of proprietary and open-source software. Specifically, the Fellow will employ the proprietary Vienna Ab-initio Simulation Package (VASP), and the open-source QuantumEspresso software suite. Given the widespread use of these codes by materials scientists in the public and private sectors, persistent open-access dissemination of raw input and output files associated with these codes is expected to be sufficient in the first instance to support the FAIRness of data associated with WP1 and WP2 (cf. Sec. 2). The primary comsumers of this data are expected to be theoretical and experimental materials scientists in the public and private sectors, who are respectively interested in reproducing and/or building upon the work undertaken by the Fellow, or in using the Fellow’s work in support of the interpretation of materials growth and characterisation.

WP3 of the Action centres on the development and application by the Fellow of a suite of software implementing numerical simulations of the electronic and optical properties of semiconductor crystal phase heterostructures based on the **k.p** method. A key deliverable of the Action is the release by the Fellow of this suite of codes as free, open-source software. It is intended that this software library – which has been designated the working name *kpymod* during its ongoing development – will take the form of a Python library (cf. Sec. 2). The primary consumers of the *kpymod* library are expected to be

theoretical and experimental materials scientists in the public and private sectors, who are interested in performing computational analysis of semiconductor heterostructures for the respective purposes of academic research into fundamental semiconductor physics and commercial development of semiconductor optoelectronic devices.

2. FAIR data

Recent developments in computational materials science have placed significant emphasis on informatics, focusing on promoting scientific reproducibility via data preservation and standardisation, by ensuring that data generated by materials simulations are made FAIR [1]. Supplementary to the adoption of rigorous procedures to ensure the security and preservation of all data associated with the Action (cf. Sec. 3), the Fellow will exploit these developments to render FAIR: (i) electronic data related to technical publications, and (ii) software developed during the Action.

In principle, it is expected that all data and software associated with all published work can be made fully open-access and FAIR by default. However, in some instances, limitations related to the dissemination of data or software used in or generated/developed during the Action either (i) already exist, or (ii) will be imposed. Instance (i) pertains to the fact that, while the Partnership Agreement between the EU Beneficiary and Partner Institution provides the Fellow with access to proprietary software developed at the Partner Institution for use in the programme of research and training associated with the Action, the Fellow is not endowed with the authority to unilaterally release such software without the consent and participation of the Partner Institution. Instance (ii) pertains to the generation and/or development of specific pieces of data and/or software by the Fellow: in cases where it may be judged that an immediate release of a specific dataset or piece of software would cause short-term damage – e.g. in terms of either diminished competitiveness, or jeporadised potential to protect intellectual property – an embargo will be imposed by the Fellow, with the intention to perform full, open-access release after such circumstances are judged to have been resolved.

2.1. Findable, accessible

The full dataset associated with each technical publication produced during the Action will be disseminated by creating a persistent, open-access archive on the Zenodo repository. The general-purpose nature of data storage enabled by Zenodo will be exploited to disseminate all raw data files associated with each publication, including relevant software required to reproduce the results and/or figures presented in the associated technical publication, as well as a README file that (i) describes the layout/contents of the data, and (ii) provides instructions regarding the reproduction of any data and/or figures described in the associated technical publication. The assignment of a unique, persistent Digital Object Identifier (DOI) to each Zenodo dataset will directly support data findability.

The Fellow will employ GitHub repositories to enable version control, persistence and open access to software developed during the Action. While such software remains under development by the Fellow and/or it is judged by the Fellow to contain underlying technical knowledge or intellectual property that respectively merits separate publication or protection, the associated GitHub repository will be retained in a private mode that restricts access to the Fellow and any contributing collaborators. The publication and/or protection of associated intellectual property will be accompanied by the conversion of the associated GitHub repository to full public access. Any GitHub repositories made public during the duration of the Action will include documentation describing the installation/use of the software.

Findability will be promoted by including a Data Access Statement in each technical publication that provides explicitly linked citations to the associated Zenodo dataset and/or GitHub repository. Using a combination of Zenodo and GitHub to respectively make datasets FAIR and to disseminate version-controlled software, readers of any publication associated with the Action will have direct, open and persistent access to the data and software associated with the Fellow’s published work. In instances where software developed during the Action is released publicly via GitHub, this release will proceed under the terms of the MIT Licence in order to promote accessibility and reusability for academic and/or commercial purposes [3].

As described above, in some instances it will not be possible to openly release proprietary software utilised in the Action. In cases where underlying software cannot be made open-access, the Zenodo

dataset associated with any technical publication(s) that employ such software will nonetheless contain all of the raw output data produced by that software – including any documentation pertinent to the formatting, interpretation and/or processing of that data – so that interested readers and/or stakeholders have full access to the data generated and/or employed by the Fellow.

2.2. Interoperable, reusable

To promote broad reusability, all open-access datasets shared via the Zenodo archive will be endowed with one of the “CC-BY” family of Creative Commons “Attribution” licences [4]. This choice has been made by the Fellow on that basis that this family licences (i) allows for flexibility in the allocation of rights and responsibilities to the end-user, and (ii) in all instances mandates that a linked reference to the original dataset is provided, with this mandated attribution intended to ensure findability and promote reusability of datasets employed in any subsequent and/or derivative work performed by the Fellow, or by other individuals or organisations. The specific choice of CC-BY licence to be applied to each dataset – e.g. whether to forbid derivative work (by applying the “ND” clause) or commerical use (by applying the “NC” clause), etc [4]. – will be decided by the Fellow on a case-by-case basis, in consultation with their Supervisors at the EU Beneficiary and Partner Institution, and with the intention of allowing the broadest possible scope for data reusability without jeopardising the potential protection or exploitation of any associated intellectual property.

As described above, WP1 and WP2 primarily consist of post-processing of data obtained from first principles calculations based on DFT, using the VASP and QuantumEspresso software. This will proceed via a combination of (i) proprietary software provided by the Partner Institution, and (ii) software developed by the Fellow. Here, reusability and interoperability of data produced by VASP and QuantumEspresso calculations is ensured via the existence of free, open-source software libraries: specifically, the Atomic Simulation environment (ASE) [5] or Python Materials Genomics (Pymatgen) [6]. These software libraries ensure data resuability by providing standardised functionality for processing and analysis of input and output files, and further ensure interoperability by allowing (i) automated conversion of input files associated with specific codes to input files associated with a wide variety of other materials software packages , and (ii) standardised post-processing of output files obtained from calculations performed using a variety of software [5,6]. It is therefore expected that open-access release of raw input/output data associated with DFT calculations performed using VASP or QuantumEspresso – via Zenodo archives associated with individual technical publications – will be sufficient to enable end-users to exploit the full scope of reusability and interoperability that is currently available to the computational materials science community.

WP3 consists of applying empirical **k.p**-based band structure models (developed in WP1) to establish numerical simulations of the electronic and optical properties of novel semiconductor heterostructures. This will involve the development by the Fellow of a suite of simulation software – working name *kpymod* (cf. Sec. 1) – which is intended to be released as free, open-source software. Reusability will be ensured by releasing the software under the terms of the MIT Licence, to allow for future academic and/or commercial use and development [3]. Interoperability of the software will be ensured via the development of an overarching Python interface for the library, which will be installable via the Package Installer for Python (pip), with documentation and version control respectively hosted on and provided via GitHub. Presenting end-users with a standard Python interface will allow to define and run simulations via scripting, thereby circumventing the requirement to define a bespoke input and output file formats. The ability to include *kpymod* library functions in Python scripts will allow for automatic interoperability with existing Python libraries, allowing the latter to be used to readily enable bespoke pre- and/or post-processing of *kpymod* input and/or output data.

While several metadata standards remain under development, there do not currently exist widely-accepted metadata standards in the areas of computational materials science relevant to the Action [2]. On the basis that the generation of appropriate metadata plays a valuable role in ensuring the FAIRness of data and software, the Fellow will review related developments continuously throughout the Action, with the intention of applying any appropriate emergent metadata standards to the data generated and software developed throughout the duration of the Action.

3. Data security

3.1. Overarching data preservation and security

Two primary layers of security will be applied to the Fellow’s data throughout and beyond the duration of the Action, providing soft and hard back-up copies of all data and software related to, and employed, generated and/or developed during the Action. These data security strategies are intended to provide robust defence against potential data loss, to mitigate the consequences of any unintentional data loss, and to guarantee persistent and remote access to the Fellow’s data.

Firstly, through their Visiting Research Fellow affiliation at the Partner Institution during the Outgoing Phase of the Action, the Fellow has access to unlimited data storage via Box (https://ucsb.app.box.com). Additionally, through their affiliation with the EU Beneficiary as a MSCA Fellow, the Fellow has access to two data storage platforms: (i) up to 5 TB of data storage via Microsoft OneDrive, associated with a TNI-UCC research staff email account (https://uccireland-my.sharepoint.com), and (ii) unlimited data storage via Google Drive, associated with a UCC academic email account (https://drive.google.com/a/ucc.ie). Access to the Box (UCSB) and OneDrive (TNI-UCC) accounts is time-limited, being active only for the duration of affiliation and/or employment of the Fellow. However, access to the Google Drive (UCC) account is persistent and perpetual. As such, the institutional UCC Google Drive account will be employed as the primary platform to ensure data security both throughout and beyond the duration of the Action, with Box and OneDrive employed to provide additional data security during the duration of the Action. The Fellow will employ the open-source, cross-platform software Rclone to push comprehensive weekly data back-ups to Google Drive, as well as secondary back-ups to Box and OneDrive. The functionality of Rclone allows to automatically maintain synced remote copies of files across multiple cloud-based platforms and operating systems, providing parallelised data transfer with built-in preservation of timestamps and explicit checksum verification [7]. Using Rclone, the Fellow can maintain regularly synced back-ups of their entire filesystem across several cloud platforms simultaneously, while also having the ability to download synced copies of their files to local computing hardware – i.e. any laptop or desktop computer having internet connectivity and a functioning Rclone installation, and running a Unix/Linux-, Windows- or macOS-based operating system.

Secondly, the Fellow retains full access to TNI-UCC computing systems via an encrypted virtual private network (VPN) during the Outgoing Phase of the Action. To provide additional data security, the Fellow will transfer regular synced filesystem back-ups to TNI-UCC’s networked file system. TNI-UCC’s IT Department runs a rigorous security protocol to protect against data loss, consisting of both soft and hard data back-ups. Automated soft copies of the networked file system are captured every 24 hours, and preserved for a peroid of up to one month after capture. At the end of each calendar month, manual hard-copy back-ups are made by writing the captured end-of-month soft filesystem copies to magnetic tape, with these magnetic tapes then stored in a secure, climate-controlled environment to provide long-term access to data back-ups in the case of unexpected local data loss.

3.2. Preservation and security of software and written work

All software developed by the Fellow throughout the duration of the Action will be secured via use of the git version-control software, in conjunction with data and software hosting provided by the GitHub platform. This approach will not only provide straightforward cloud-based back-ups of software and associated documentation, but also rigorous version control to ensure software integrity.

Written work associated with the Action – including research notes and reports, software documentation, technical publications, etc. – will be generated, version-controlled, backed up and shared with collaborators via the use of cloud-based platforms. In particular, it is expected that all technical written work associated with the Action will be typeset using the LaTeX markup language, in which case the Overleaf platform will be employed (https://overleaf.com). This use of cloud-based editors is intented to provide persistence and security of all documentation associated with the Action, while in the case of technical documentation and publications Overleaf’s built-in GitHub integration will be exploited to generate a version-controlled repository associated with each individual document.

3.3. Preservation and dissemination of data related to technical publications

All data associated technical publications arising directly from, or building upon, work carried out during the Action will be made persistently secure, FAIR and openly accessible via the creation of archives in certified data repositories. It is intended that this will primarily take the form of the creation of a distinct Zenodo archive associated with each individual publication (cf. Sec. 2).

4. Allocation of resources

All data storage, security, preservation and dissemination strategies associated with the Fellowship are accessible at no financial cost to the Action. All such strategies described above require resources that are either provided to the Fellow by the EU Beneficiary and/or Partner Institution (where access has already been granted), or consist of open-access and/or open-source software, platforms and/or archives (where the Fellow is presented with no barriers to entry, or limits to continuous usage).

5. Ethical and legal aspects

There are no ethical or legal issues that impact the generation, storage or dissemination of data and/or software generated and/or developed during this specific Action.

References

[1] M. Scheffler, M. Aeschlimann, M. Albrecht, T. Bereau, H.-J. Bungartz, C. Felser, M. Greiner, A. Gross, C. T. Koch, K. Kremer, W. E. Nagel, M. Scheidgen, C. Woll, and C. Draxl, “FAIR data enabling new horizons for materials research”, *Nature* **604**, 635 (2022). DOI: 10.1038/s41586-022-04501-x

[2] “The EMMC Roadmap for Materials Modelling and Digitalisation of the Materials Sciences”, N. Adamovic, J. Friis, G. Goldbeck, A. Hashiborn, K. Hermansson, D. Hristova-Bogaerds, R. Koopmans, and E. Wimmer (European Materials Modelling Council, 2020). DOI: 10.5281/zenodo.4272033

[3] “White Paper for Standards of Modelling Software Development”, European Materials Modelling Council (2018)

[4] T. Rathmann, “Licenses for Research Data” (2018). DOI: 10.5281/zenodo.1463156

[5] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, K. W. Jacobsen, “The Atomic Simulation Environment – A Python Library for Working With Atoms”, *J. Phys.: Condens. Matter* **29**, 273002 (2017). DOI: 10.1088/1361-648X/aa680e

[6] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, “Python Materials Genomics (Pymatgen): a robust, open-source Python library for materials analysis”, *Comput. Mater. Sci.* **68**, 314 (2013). DOI: 10.1016/j.commatsci.2012.10.028

[7] S. Rivera, J. Griffieon, Z. Fei, M. Hayashida, P. Shi, B. Chitre, J. Chappell, Y. Song, L. Pike, C. Carpenter, and H. Nasir, “Navigating the Unexpected Realities of Big Data Transfers in a Cloud-based World”, *Proceedings of the Practice and Experience on Advanced Research Computing* (2018). DOI: 10.1145/3219104.3229276

Version history

Date	Version no.	Author	Notes
13/05/2022	0.1	C. A. Broderick	Initial draft
03/06/2022	1.0	C. A. Broderick	Submitted version