ANR DYRE-COMB DMP

GENERAL INFORMATION

Administrative information

Acronym: DYRE-COMB
Decision code: Project **ANR-22-CE49-0016**
Title: DYnamo REgimes dependence on the heterogeneous COre-Mantle Boundary heat flux
Coordinator's last name: Amit
Coordinator's first name: Hagay
Coordinator's email: Hagay.Amit@univ-nantes.fr
Coordinator's Affiliation: Laboratoire de Planétologie et Géosciences, Nantes, France.
DMP contact person: Aude Chambodut (aude@unistra.fr, Institut Terre et Environnement de Strasbourg, France)

DMP version	*Date*	*Contributors, inputs and changes*
Initial V0.0	07/09/2023	H. Amit, G. Choblet – Initial draft version following first 6-months of project
V0.1	07 to 25/09/2023	H. Amit, G. Choblet, F. Deschamps, E. Thébault, A. Chambodut – references / explanation of code, data and metadata / data repository
V0.2	25/09/2023	H. Amit, A. Chambodut – Final version for submission of DMP to ANR at (Start-Date + 6 months)

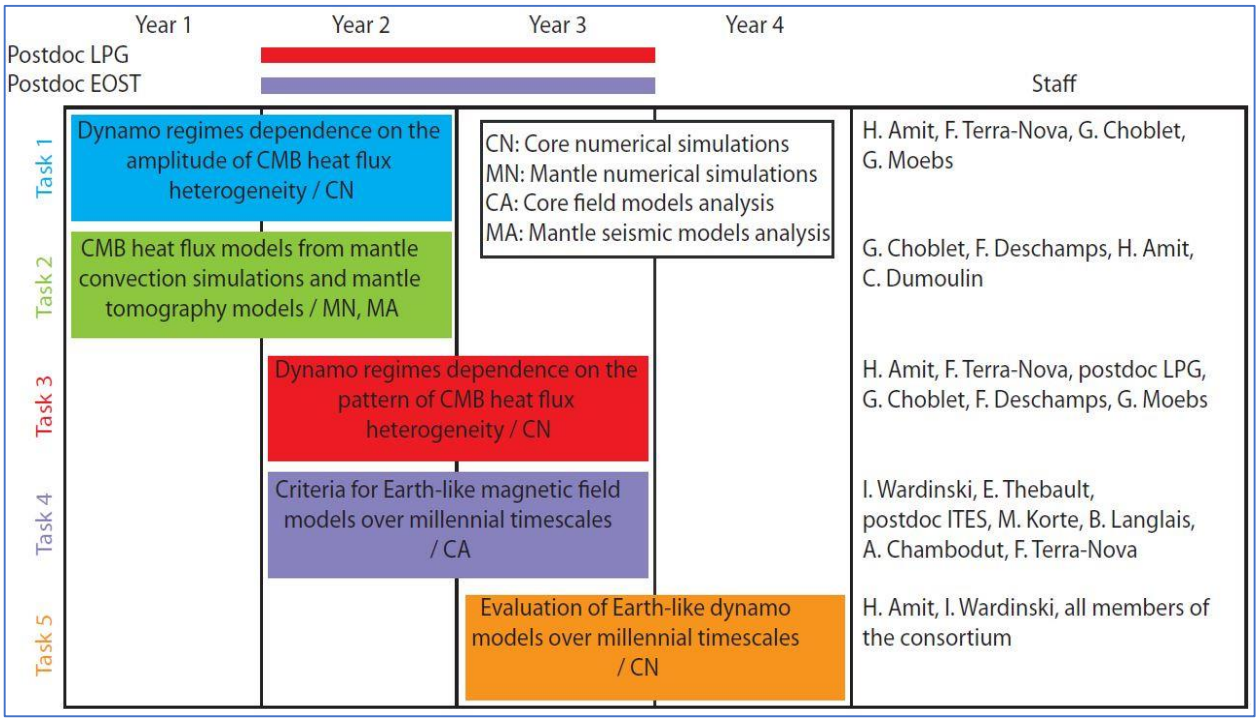

Figure 1: The project's tasks, work schedule, organization and staff.

1. DATA DESCRIPTION AND COLLECTION OR RE-USE OF EXISTING DATA

1a. How will new data be collected or produced and/or how will existing data be re-used?

Guidance:

- Explain which methodologies or software will be used if new data are collected or produced.
- State any constraints on re-use of existing data if there are any.

- Explain how data provenance will be documented.
- Briefly state the reasons if the re-use of any existing data sources has been considered but discarded.

The main data that will be produced by the project are **simulation-based models by two codes**:

- the core convection code MagIC (Wicht, 2002; https://magic-sph.github.io/) [tasks 1, 3 and 5] and
- the mantle convection code StagYY (Tackley, 2008; https://github.com/ptackley/StagYY/tree/master/doc) [task 2].

Additional used data-products include existing **observation-based models**, seismic tomography models [task 2] and geomagnetic field models [task 4], and their analysis.

No specific methodology will be used to collect any new data or data-products as there are already existing codes for the simulations. In addition, no specific methodology will be used to extract existing models as these are documented in peer-reviewed journals that are widely used by the community. However, some short codes will be designed to analyse the existing models of both tomography and magnetic field, e.g. to extract timeseries of some morphological properties of these models.

- Wicht, J., 2002. Inner-core conductivity in numerical dynamo simulations. *Phys. Earth Planet. Inter.*, 132, 281-302. doi:10.1016/S0031-9201(02)00078-X
- Tackley, P.J., 2008. Modelling compressible mantle convection with large viscosity contrasts in a three-dimensional spherical shell using the yin-yang grid. *Phys. Earth Planet. Inter.*, 171, 7-18. doi:10.1016/j.pepi.2008.08.005

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

Guidance:

- Give details on the kind of data: for example numeric (databases, spreadsheets), textual (documents), image, audio, video, and/or mixed media.
- Give details on the data format: the way in which the data is encoded for storage, often reflected by the filename extension (for example pdf, xls, doc, txt, or rdf).
- Justify the use of certain formats. For example, decisions may be based on staff expertise within the host organisation, a preference for open formats, standards accepted by data repositories, widespread usage within the research community, or on the software or equipment that will be used.
- Give preference to open and standard formats as they facilitate sharing and long-term re-use of data (several repositories provide lists of such 'preferred formats').
- Give details on the volumes (they can be expressed in storage space required (bytes), and/or in numbers of objects, files, rows, and columns).

Code	Data type	Format	Justification	Typical size	# runs or models
MagIC	Time-dependent 3D simulation-based models	Non-formatted binary	Very heavy files	~1 Gigabyte	~100 runs
StagYY	Time-dependent 3D simulation-based models	Non-formatted binary	Very heavy files	~1 Gigabyte	~10 runs
Analysis of geomagnetic field observation-based models	Various types of analysis output	Standard text	Direct use in visualization codes	~1 Megabyte	~20 models
Analysis of seismic tomography observation-based models	Various types of analysis output	Standard text	Direct use in visualization codes	~1 Megabyte	~5 models

The formats are conventional widely-used within the scientific communities that are studying fluid dynamics in deep Earth envelopes.

2. DOCUMENTATION AND DATA QUALITY

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

Guidance:

- Indicate which metadata will be provided to help others identify and discover the data.
- Indicate which metadata standards (for example DDI, TEI, EML, MARC, CMDI) will be used. Use community metadata standards where these are in place.
- Indicate how the data will be organised during the project, mentioning for example conventions, version control, and folder structures. Consistent, well-ordered research data will be easier to find, understand, and re-use.
- Consider what other documentation is needed to enable re-use. This may include information on the methodology used to collect the data, analytical and procedural information, definitions of variables, units of measurement, and so on.
- Consider how this information will be captured and where it will be recorded for example in a database with links to each item, a 'readme' text file, file headers, code books, or lab notebooks.

For each one of the two simulations, core convection (MagIC) and mantle convection (StagYY), two types of metadata exist:

1. The **input files** that define the physical problem (geometry of the fluid shell, non-dimensional numbers, physical approximations, boundary conditions) and the choice numerical resolution (grid size in physical and spectral domains, temporal stepping parameters, calculation scheme, parallel computing characteristics, nature of output and diagnostic files).
2. **List of simulations** that were performed.

No standard exists for these metadata. However, the core convection code MagIC and the mantle convection code StagYY are open source and are freely available with published documentations that describe the type 1 metadata, which guarantees a simple access for reproducing the results. Concerning the list of simulations (metadata type 2), we will provide a full list of the obtained simulation-based models of each code.

For the geomagnetic field models, metadata will consist of the characterization of their morphologies [task 4]. A list of these morphological characterizations will be provided for each existing archeomagnetic and historical field model. **For the seismic tomography models**, metadata will consist of the inferred CMB heat flux model [task 2]. A list of CMB heat flux models will be provided for each existing seismic tomography model.

2b. What data quality control measures will be used?

Guidance:

- *Explain how the consistency and quality of data collection will be controlled and documented. This may include processes such as calibration, repeated samples or measurements, standardised data capture, data entry validation, peer review of data, or representation with controlled vocabularies.*

Proper spatial resolution in both core and mantle convection simulation-based models will be controlled by ensuring that the energies spectra exhibit a sufficiently large peak to tail drop. Proper temporal resolution will be ensured by a well-behaved force balance, e.g. in the core convection simulations the pressure gradient force must be equilibrated by some other force at all scales. Obviously proper dynamo simulation must not have a decaying magnetic energy. Similar controls will be applied for the mantle convection simulations, i.e. verify that global scalars (mean temperature, rms velocity, top and bottom heat flux) oscillate about a steady state rather than drift and that the radial energy balance precisely holds.

Proper analysis of observation-based geomagnetic field and mantle tomography models will attribute error bars that will account for the data uncertainties as reported by the authors of the models.

3. STORAGE AND BACKUP DURING THE RESEARCH PROCESS

3a. How will data and metadata be stored and backed up during the research?

Guidance:

- Describe where the data will be stored and backed up during research activities and how often the backup will be performed. It is recommended to store data in least at two separate locations.
- Give preference to the use of robust, managed storage with automatic backup, such as provided by IT support services of the home institution. Storing data on laptops, standalone hard drives, or external storage devices such as USB sticks is not recommended.

The code MagIC will be run on the regional computation facility GLICID (https://www.glicid.fr/) used by LPG Nantes. The source code, input file and all output simulation-based models will be stored in the regional computation facility GLICID at first stage. Then all files will be backed up locally at LPG (hard disks, personal computers) and remotely at ITES (seafile). We are aware that this is not satisfactory, we will work on an improved storage scheme for future DMP versions. For the short term, GLICID provides backed up storage for the output data of the MagIC models.

The code StagYY will be run in Academia Sinica Taipei. The source code is stored on the IES linux cluster PlanetB612, which is regularly backed up, as well as on individual workstations and external hard drives. After simulations finished, simulation-based models will be stored on at least three different storage systems, including the research group's NAS and additional external hard disks.

The metadata of the two convection codes will be stored on personal posts. Far simpler codes for the analysis of geomagnetic field models and seismic tomography models will be run on personal posts.

The output of observation-based models is expected to be light, so no particular storage strategy is required.

3b. How will data security and protection of sensitive data be taken care during the research

Guidance:

- *Explain how the data will be recovered in the event of an incident.*
- *Explain who will have access to the data during the research and how access to data is controlled, especially in collaborative partnerships.*
- *Consider data protection, particularly if your data is sensitive for example containing personal data, politically sensitive information, or trade secrets. Describe the main risks and how these will be managed.*
- *Explain which institutional data protection policies are in place.*

All the data is accessible to all project's participants. Data exchange is performed via shared spaces such as UNcloud of the Nantes University. The data produced by the DYRE-COMB project is not sensitive.

4. LEGAL AND ETHICAL REQUIREMENTS, CODE OF CONDUCT

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

Guidance:
Ensure that when dealing with personal data protection laws (for example GDPR) are complied with:

- *Gain informed consent for preservation and/or sharing of personal data.*
- *Consider anonymisation of personal data for preservation and/or sharing (truly anonymous data are no longer considered personal data).*
- *Consider pseudonymisation of personal data (the main difference with anonymisation is that pseudonymisation is reversible).*
- *Consider encryption which is seen as a special case of pseudonymisation (the encryption key must be stored separately from the data, for instance by a trusted third party).*
- *Explain whether there is a managed access procedure in place for authorised users of personal data.*

Data used and produced are simulation-based models of convection in Earth's mantle and core as well as observation-based models of the geomagnetic field on the core-mantle boundary and the seismic anomalies at the lower mantle. No legal or ethical aspects are involved, no commercial use is envisioned.

Metadata and forthcoming publications or communications will refer to persons, i.e. authors and owners of the data and codes used. However, these information pieces will strictly follow the GDPR (EU General Data Protection Regulation 2016/679, https://gdpr-info.eu/) and constitute the acknowledgement and citation of previous peer and colleagues' works. Apart from citation and participant's personal authentication to the website/wiki/cloud of the project, no personal data is involved (collected, used or transferred) in this project.

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

Guidance:

- *Explain who will be the owner of the data, meaning who will have the rights to control access: Explain what access conditions will apply to the data? Will the data be openly accessible, or will there be access restrictions? In the latter case, which? Consider the use of data access and re-use licenses.*
- *Make sure to cover these matters of rights to control access to data for multi-partner projects and multiple data owners, in the consortium agreement.*

- *Indicate whether intellectual property rights (for example Database Directive, sui generis rights) are affected. If so, explain which and how will they be dealt with.*
- *Indicate whether there are any restrictions on the re-use of third-party data.*

All the data produced in the project will be subject to French law (*) under the supervision of CNRS and in the respect of rules laid down by *Agence National de la Recherche* that funds the present project.

Licence onto simulation-based models and various analyses of observation-based models will be under Creative Commons Attribution 4.0 International Public License (https://creativecommons.org/licenses/by/4.0/legalcode)

(*) French Law n° 2016-1321 of 7 October 2016 for a "Digital Republic" (*République numérique*): https://www.legifrance.gouv.fr/jorf/article_jo/JORFARTI000033202841

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

Guidance:

- *Consider whether ethical issues can affect how data are stored and transferred, who can see or use them, and how long they are kept. Demonstrate awareness of these aspects and respective planning.*
- *Follow the national and international codes of conducts and institutional ethical guidelines, and check if ethical review (for example by an ethics committee) is required for data collection in the research project.*

There are no ethical issues associated with this project. No ethical review is required for the used and created data collection in the research project.

5. DATA SHARING AND LONG-TERM PRESERVATION

5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Guidance:

- *Explain how the data will be discoverable and shared (for example by deposit in a trustworthy data repository, indexed in a catalogue, use of a secure data service, direct handling of data requests, or use of another mechanism).*
- *Outline the plan for data preservation and give information on how long the data will be retained.*
- *Explain when the data will be made available. Indicate the expected timely release. Explain whether exclusive use of the data will be claimed and if so, why and for how long. Indicate whether data sharing will be postponed or restricted for example to publish, protect intellectual property, or seek patents.*
- *Indicate who will be able to use the data. If it is necessary to restrict access to certain communities or to apply a data sharing agreement, explain how and why. Explain what action will be taken to overcome or to minimise restrictions.*

At the beginning of the project all the codes and data products are by nature already properly referenced and available for the international scientific community. During the project, all produced data and data products will be reported in peer-reviewed publications. For publication (just after acceptation of article), selected(*) results will be shared and ingested into a thematic or generic data repository with long-term preservation policy.

(*) Due to the huge size of the simulations output.

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

Guidance:

- *Indicate what data must be retained or destroyed for contractual, legal, or regulatory purposes.*
- *Indicate how it will be decided what data to keep. Describe the data to be preserved long-term.*
- *Explain the foreseeable research uses (and/or users) for the data.*
- *Indicate where the data will be deposited. If no established repository is proposed, demonstrate in the data management plan that the data can be curated effectively beyond the lifetime of the grant. It is recommended to demonstrate that the repositories policies and procedures (including any metadata standards, and costs involved) have been checked.*

We are currently considering what fraction of the data output published in peer-reviewed journals is required for archiving and long-term preservation in the objective of research reproducibility. Due to the very large size of the output files of the convection codes, only part of the output will be preserved. For each simulation, we will preserve a full starting point (containing all fields in 3D), full end point and an input file metadata which allows reproduction of all the results. In addition, the temporal evolution of some key diagnostics (e.g. energies, dipole properties for the dynamo code, radial profiles, zonal profiles) will also be preserved. This aspect will be elaborated in a future version of the DMP as a collegial work.

As no dedicated data repository exists for dynamo simulations, at international, European or national level, the generic *RechercheDataGouv* data repository (https://recherche.data.gouv.fr/en) is currently planned to be used. Indeed, its subspace CNRS handled by CNRS-DDOR and subsubspace INSU appears to be a reasonable choice. *RechercheDataGouv* data repository allows to properly reference all involved laboratories, to clearly show all affiliations of project's participants (such as CNRS-INSU OSU, universities, or any other supervisory body) and to acknowledge funders (ANR).

It has to be noted that discussions are currently ongoing (end of 2023) between the research community of geodynamo modellers and the e-infrastructure Data Terra (data hub ForM@Ter for solid earth) in order to possibly define a dedicated data repository. Obviously if a national data repository will especially be designed for geodynamo models and outputs, the current project will take advantage of it.

5c. What methods or software tools are needed to access and use data?

Guidance:

- *Indicate whether potential users need specific tools to access and (re-)use the data. Consider the sustainability of software needed for accessing the data.*
- *Indicate whether data will be shared via a repository, requests handled directly, or whether another mechanism will be used?*

No specific methods or software tools are needed to access and use data.

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

Guidance:

- *Explain how the data might be re-used in other contexts. Persistent identifiers should be applied so that data can be reliably and efficiently located and referred to. Persistent identifiers also help to track citations and re-use.*
- *Indicate whether a persistent identifier for the data will be pursued. Typically, a trustworthy, long-term repository will provide a persistent identifier.*

The codes MagIC and StagYY are publicly available on the platform GitHub and referenced in peer-reviewed publications (see Section 1 of the present DMP).

The data will become available for sharing upon publication of articles. All the data and the metadata that will be used for the writing of an article will become accessible on the chosen data repository (national generic data platform *RechercheDataGouv* or thematic data repository of the national e-infrastructure Data Terra), which will allow DOI minting to a deposit.

Publications (in peer-reviewed journals) are automatically getting a DOI. They will also be made available through HAL Open Science (https://hal.science/) as required by CNRS.

6. DATA MANAGEMENT RESPONSIBILITIES AND RESOURCES

6a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

Guidance:

- *Outline the roles and responsibilities for data management/stewardship activities for example data capture, metadata production, data quality, storage and backup, data archiving, and data sharing. Name responsible individual(s) where possible.*
- *For collaborative projects, explain the co-ordination of data management responsibilities across partners.*
- *Indicate who is responsible for implementing the DMP, and for ensuring it is reviewed and, if necessary, revised.*
- *Consider regular updates of the DMP.*

The master co-ordinator of data management responsibilities across partners will be Aude Chambodut. Each partner will be responsible of the data used in the assigned task and related metadata documentation. The DMP will be regularly updated after each publication.

6b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

Guidance:

- *Explain how the necessary resources (for example time) to prepare the data for sharing/preservation (data curation) have been costed in. Carefully consider and justify any resources needed to deliver the data. These may include storage costs, hardware, staff time, costs of preparing data for deposit, and repository charges.*
- *Indicate whether additional resources will be needed to prepare data for deposit or to meet any charges from data repositories. If yes, explain how much is needed and how such costs will be covered*

Storage and saving of data and metadata is under the responsibility of each project participant that produces these data and metadata. The time associated with these tasks is modest and is included in the research project. Possible costs are covered by the DYRE-COMB budget.

Data sorting for long-term preservation will result from discussions among all the participants of the project under the responsibility of the leaders of each partner: Hagay Amit (LPG Nantes), Ingo Wardinski (ITES Strasbourg) and Erwan Thébault (LMV Clermont-Ferrand) with the coordination of Aude Chambodut.

Each partner of the present project are members of *Observatoires des Sciences de l'Univers* (OSU):

- EOST- *Ecole et Observatoire des Science de la Terre*, Strasbourg,
- OPGC - *Observatoire de Physique du Globe de Clermont-Ferrand*, or
- OSUNA – *Observatoire des Sciences de l'Univers de Nantes Atlantique*.

OSU are French structures that implement national observation services and initiatives over the long term, and promote interdisciplinary research. In particular, they facilitate access to major research infrastructures and experimental or digital platforms of regional or national interest. OSU missions encompass the data management aspects in particular FAIR data and TRUST data repository principles. Some OSU developed data repositories for all data handled in their perimeter from observation (national services of observation, i.e.: INSU SNO), research (laboratories, i.e.: CNRS UMR) and teaching (training datasets).

AGENCE NATIONALE DE LA RECHERCHE
ANR
www.anr.fr
@agencerecherche ANR