# General Framework for Creative Artificial Intelligence

*A Data Management Plan created using DMPonline.be*

**Creators:** n.n. n.n., Thomas Winters

**Affiliation:** KU Leuven (KUL)

**Funder:** KU Leuven (KUL)

**Template:** KU Leuven BOF-IOF

**Project Administrator:** n.n. n.n.

**Grant number** / **URL:** PDMT2/23/050

**ID:** 205185

**Start date:** 06-12-2023

**End date:** 31-10-2024

**Project abstract:**

In recent years, the field of creative artificial intelligence has seen remarkable growth, thanks to the advancements in generative models like ChatGPT, DALL-E, and others. Two of the most significant challenges in this domain are control over creative AI models and the need for a general framework for integrating multiple generative models. Current implementations often need more fine-grained control and transparency, making them less suitable for integration into professional creative processes. Integrating different generative models is usually done ad hoc and lacks a formal approach for reasoning and control. This postdoctoral application proposes to address these challenges by developing a modeling language for creative AI called FlowState. This language will create new conceptualizations to formalize the integration of generative models. Creating inference mechanisms for inspecting and constraining creative AI models helps make their generations adhere to properties desired by the user. The project will also explore the feasibility and effectiveness of FlowState in various creative domains, including text and image generation for poetry, humor, comic books and animation. The successful conceptualization and implementation of FlowState offer a more coherent, controllable, efficient, and transparent approach to integrating various generative models and potentially lead to new AI-augmented creativity possibilities.

**Last modified:** 04-03-2024

## General Framework for Creative Artificial Intelligence

**Research Data Summary**

**List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.**

| Dataset name / ID | Description | New or reuse | Digital or Physical data | Data Type | File format | Data volume | Physical volume |
|---|---|---|---|---|---|---|---|
| | | *Indicate:* ***N****(ew data) or* ***E****(xisting data)* | Indicate: **D**(igital) or **P**(hysical) | Indicate: **A**udiovisual **I**mages **S**ound **N**umerical **T**extual **M**odel **SO**ftware Other (specify) | | Indicate: <1GB <100GB <1TB <5TB >5TB NA | |
| WP2.1 | Generative probability dataset | N | D | T, N | Markup language | <1GB | |
| WP2.2 | Generative space inferences | N | D | T, N | Markup language / Spreadsheets | <1GB | |
| WP3.1 | Constrained generation analytics | N | D | T, N | Markup language / Spreadsheets | <1GB | |
| WP3.2 | Neural constraints performance | N | D | N | Markup language / Spreadsheets | <1GB | |
| WP4.1 | Induction examples dataset | E | D | T | Markup language | <1GB | |
| WP4.2 | Text-to-Generator data | N | D | T | Markup language | <1GB | |
| WP5 | Artefact evaluations | N | D | T, I, S, N | Markup language / Spreadsheets | <1GB | |
| Scripts | Scripts used to build models & analyse | N | D | SO | python | <1GB | |

**If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:**

We will use existing datasets for training examples, but which ones will depend on the actual papers and are still to be decided.

**Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.**

- No

**Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).**

- No

**Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.**

- No

**Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material or Data transfer agreements, Research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.**

- No

**Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.**

- No

#### Documentation and Metadata

**Clearly describe what approach will be followed to capture the accompanying information necessary to keepdata understandable and usable, for yourself and others, now and in the future (e.g. in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, codebook.tsv etc. where this information is recorded).**

The code, models and datasets used in our published papers will be made available on open-source platforms (GitHub, Huggingface). These have README files to describe the data and code.
Long term, a copy of all the necessary programs, environments, models and data is also always saved on the DTAI ML Archive, which is a NetApp storage system that is managed by the CS Department.

**Will a metadata standard be used to make it easier tofind and reuse the data?**
**If so, please specify which metadata standard will be used.**

**If not, please specify which metadata will be created to make the data easier to find and reuse.**

- Yes

Yes, we use json and csv for our data and open standards for saving tensors for our models.

#### Data Storage & Back-up during the Research Project

**Where will the data be stored?**

- Other (specify below)
- Large Volume Storage

- Personal network drive (I-drive)
- OneDrive (KU Leuven)

During the research, the programs are stored on personal laptops and synchronized with code platforms (e.g. Github). Reports, data and other results are stored on online cloud drives from KU Leuven.

**How will the data be backed up?**

- Standard back-up provided by KU Leuven ICTS for my storage solution
- Personal back-ups I make (specify below)

Back-ups are synchronized online using the code platforms and online cloud drives (KU Leuven OneDrive).
We archive our programs, models, code and data on the DTAI ML Archive folder, on the NetApp system managed by the computer science department.

**Is there currently sufficient storage & backup capacity during the project?**

**If no or insufficient storage or backup capacities are available, explain how this will be taken care of.**

- Yes

**How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?**

Using the rights management provided by the code platforms (GitHub and HuggingFace) and KU Leuven online cloud drive rules.

**What are the expected costs for data storage and backup during the research project? How will these costs be covered?**

Included in already existing plans from KU Leuven (free usage of code synchronization & existing online drive tools).

**Data Preservation after the end of the Research Project**

**Which data will be retained for 10 years (or longer, in agreement with other retention policies that are applicable) after the end of the project?**

**In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).**

- All data will be preserved for 10 years according to KU Leuven RDM policy

**Where will these data be archived (stored and curated for the long-term)?**

- Large Volume Storage (longterm for large volumes)
- Shared network drive (J-drive)

We archive our programs, models, code and data on the DTAI ML Archive folder, managed by the computer science department.

**What are the expected costs for data preservation during the expected retention period? How will these costs be covered?**

Costs are covered by the ML research group. Contact wannes.meert@kuleuven.be for more information about this storage.

#### Data Sharing and Reuse

**Will the data (or part of the data) be made available for reuse after/during the project?**
**Please explain per dataset or data type which data will be made available.**

- Yes, as open data

New datasets will be made available appropriate online sharing platforms (such as GitHub & HuggingFace) when they are relevant to the community, e.g. for training or analytical purposes. More detailed datasets with less relevant artefacts will be stored but only available upon request.

**If access is restricted, please specify who will be able to access the data and under what conditions.**

N/A

**Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?**

**Please explain per dataset or data type where appropriate.**

- No

**Where will the data be made available?**

**If already known, please provide a repository per dataset or data type.**

- Other data repository (specify below)

GitHub & HuggingFace

**When will the data be made available?**

- Upon publication of research results

**Which data usage licenses are you going to provide?**

**If none, please explain why.**

- CC-BY 4.0 (data)
- MIT licence (code)

**Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.**

- Yes, a PID will be added upon deposit in a data repository

**What are the expected costs for data sharing? How will these costs be covered?**

Typically free

**Responsibilities**

**Who will manage data documentation and metadata during the research project?**

Thomas Winters

**Who will manage data storage and backup during the research project?**

Thomas Winters

**Who will manage data preservation and sharing?**

Thomas Winters (short-term, during project), Wannes Meert & Luc De Raedt (long-term, after project)

**Who will update and implement this DMP?**

Thomas Winters (short-term, during project), Wannes Meert & Luc De Raedt (long-term, after project)