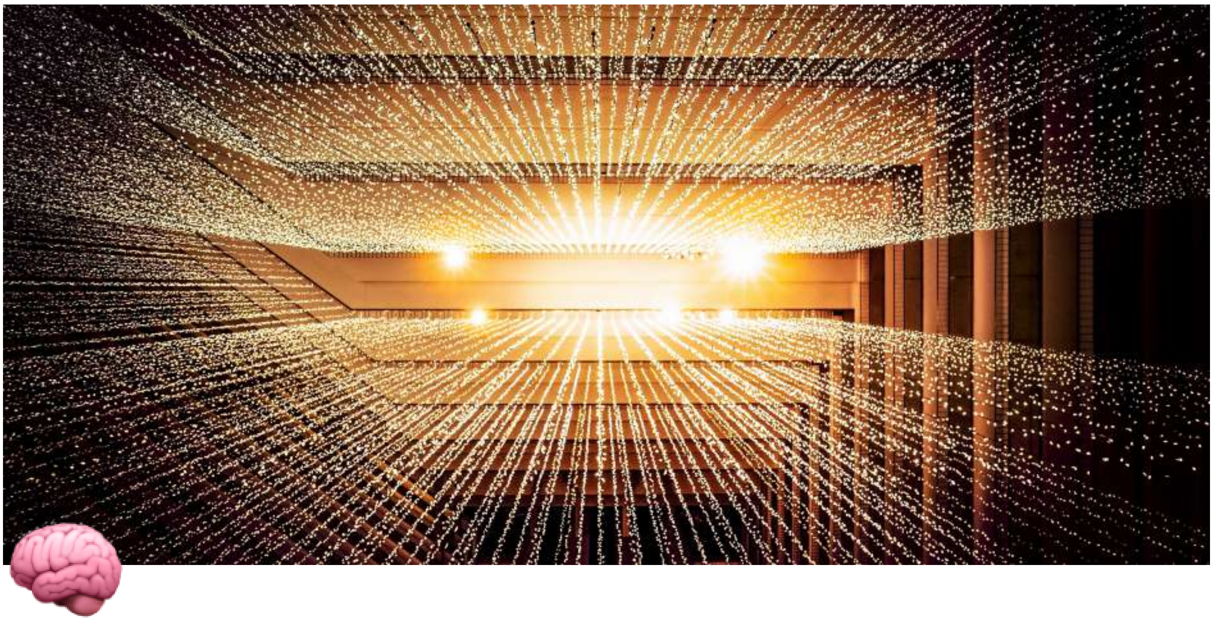

Swarm Fellowship Canvas: Data and model provenance for decentralized AI

Description

The Data and model provenance for decentralized AI Fellowship Project aims to develop a toolset for use in data and model provenance in AI applications, leveraging Swarm decentralized storage and blockchain technology, addressing critical needs in ethical AI development and regulatory compliance while showcasing the potential of Swarm and other Web3 technologies.

The key objectives are to address the need for recorded provenance of data and models in AI, ensure ethical practices and regulatory compliance in AI development, implement a system for tracking and recording data origins and transformations and utilize Web3 technologies for secure attestation and verification.

Secondary objectives are to deliver research on Swarm attractiveness to AI companies and community engagement through the Swarm Improvement Proposal (SWIP) process.

Foundation

Purpose

Why are we doing the project?

We are undertaking this project to address what we feel is a critical need for recorded provenance of data and models in AI (in short - data), ensuring ethical practice and regulatory compliance. We also want to validate and focus on that need by interviewing various stakeholders and engaging the community through the SWIP process.

Summarise in a couple of sentences and describe the context in several paragraphs.

In the landscape of AI and data management, the provenance of data and models is becoming increasingly important (see 1, 2). Ethical practices and regulatory frameworks mandate that the origins and transformations of data and models used in AI are transparently tracked and recorded (see 3). Provenance ensures accountability, integrity, and trustworthiness of data, which is essential for ethical AI applications.

Data and models with recorded provenance are inherently more valuable and reliable, making them preferable for use in AI applications. By ensuring comprehensive provenance, we enhance the value and usability of data and models, promoting trust and compliance in AI systems. Provenance and interoperability also allows for new use cases to emerge, such as crowd sourcing of foundational models, built on contributed datasets.

Swarm decentralized storage offers unique features such as immutability, self-sovereignty, and independence from any single provider or data silo. Leveraging these capabilities, along with Layer 2 (L2) solutions, we should aim to store datasets and AI models in such a secure and decentralized manner.

The project will address the provenance challenge by implementing a system (Toolkit) that tracks and records the origin and transformations of data and models as they pass through various stages and users. This solution will employ Web3 technologies, specifically Swarm storage and blockchain with smart contracts, to enable secure attestation and verification of data along the chain. Code will be open sourced and available for the ecosystem to use, building towards a foundation for interoperability.

What is the business use case for the project - how will it be maintained after the fellowship ends. Describe in a couple of paragraphs.

The business use case for this project centers on providing a toolkit using decentralized technologies that can be used for data and model provenance, a regulatory requirement with significant business benefits. This Toolkit, leveraging Swarm decentralized storage and blockchain technology, will appeal to numerous parties across industries such as healthcare, finance, and research. By being open-source, it invites community contributions, fostering continuous improvement and expansion based on collective needs.

Moreover, in the context of Datafund and its focus on AI, this Toolkit will be integral to solutions supporting decentralized AI and the Fair Data Economy. The open-source components will be actively used and maintained as part of Datafund's business initiatives, drawing in a larger ecosystem of users.

This dual approach of regulatory compliance and business integration ensures sustained interest and ongoing development, making the Toolkit a vital resource in the future data economy.

Benefits

What benefits and impacts will the project generate and how will we know the project is successful?

List general benefits in a couple of sentences and describe them in more detail in several paragraphs.

The project is expected to position Swarm and Web3 technologies as independent solutions for ensuring data and model provenance, complying with regulatory requirements and increasing the trust of and the value of the data and models. This will benefit the Web3 ecosystem, particularly decentralized storage and Swarm, while also providing significant advantages for AI developers, regulators, Swarm users and communities, and the broader Web3 community.

Detailed Benefits:

1. **Enhanced Traceability and Integrity:**
    - The project leverages Web3 technologies, including Swarm decentralised storage and blockchain, to provide robust traceability and integrity of data and models. The immutability guarantees of these technologies, combined with attestation and verification mechanisms, ensure that the provenance of data and models is securely recorded and verifiable throughout the entire chain.

2. **Regulatory Compliance:**
    - By ensuring comprehensive provenance, the project helps stakeholders comply with ethical practices and regulatory requirements. This is particularly important in the AI domain, where accountability and transparency are critical for maintaining trust and integrity.
3. **Open Source Tooling Available:**
    - Open source tooling will be available for developers to easily implement data provenance in their solution - for app developers, or for their data - for AI developers. This will make Swarm more attractive to build on, especially for AI related projects.
4. **Visibility and Use Cases:**
    - Given the burgeoning interest in AI, the project is poised to generate significant visibility around the use of decentralized storage for AI applications. This can lead to numerous potential use cases and innovations, further driving the adoption and development of Web3 technologies.
5. **Long-Term Adoption and Fair Data Economy:**
    - In the long term, the project aims to increase the adoption of decentralized storage solutions, even for smaller and niche datasets. This aligns with fair data economy practices by promoting equitable access to data and models, fostering innovation, and supporting smaller players in the data ecosystem.

People

Resources

Who will manage the project, and which skills are needed to deliver the project?

List the skills needed to do the project.

- **Project Manager**
    - **Description:** Oversees the entire project, ensuring that it stays on schedule, within budget, and meets its objectives. Manages

communication between stakeholders, coordinates team activities, and addresses other issues that arise.

- **Lead Developer and Requirements Analyst**
  - **Description:** Responsible for the overall technical direction of the project and the SWIP process. Ensures the development team adheres to specifications and best practices, provides technical guidance, solves complex coding challenges, and manages the requirements gathering and SWIP proposal process. Also handles documentation, data management, and engages with stakeholders to gather requirements.
- **Frontend Developer**
  - **Description**: Focuses on the user interface and user experience of the Platform. Implements the visual and interactive aspects of the application.
- **Backend Developer**
  - **Description**: Works on the server-side logic, database interactions, and integration with decentralized storage solutions like Swarm.
- **Blockchain Developer**
  - **Description**: Specializes in developing and integrating blockchain technology, working on smart contracts and ensuring secure and transparent transactions. Developers also provide technical support to users of the Platform.
- **Tester/Quality Assurance (QA) Specialist**
  - **Description:** Conducts testing of the Platform to identify bugs and issues. Ensures that the final product meets quality standards and performs as expected.
- **DevRel Specialist**
  - **Description**: A DevRel (Developer Relations) Specialist acts as a bridge between the development team and the external developer community. They play a crucial role in fostering adoption, engagement and a thriving developer ecosystem. One of the outlets for communication is the documentation. Engages with the community, facilitates discussions around the SWIP process, gathers feedback, and communicates updates.

Stakeholders

Who will benefit from and be affected by the project?

List all the groups and individuals affected by the project and describe in a couple of sentences how they will be affected.

AI developers will benefit from access to open source tooling for data provenance, as well as reliable and verifiable data and models, enhancing the quality and accountability of their AI systems through robust provenance tracking and secure storage solutions.

Data providers and owners will benefit from secure and transparent value chain for their contributions, with increased trust in data transactions due to accurate tracking and ethical use of their data.

End-users of AI applications will gain from increased trust and reliability in AI systems, as the provenance of data and models ensures high-quality and ethically sourced data.

Regulators will gain improved tools for ensuring compliance with ethical practices and regulatory requirements, with transparent and immutable records of data provenance aiding in monitoring and enforcement.

Swarm users and the broader Swarm community will see enhanced functionality and value from Swarm decentralized storage, driving adoption and engagement through a high-visibility, impactful use case.

The **Web3 ecosystem** will be strengthened by showcasing practical applications of decentralized technologies, fostering innovation and broader adoption within the ecosystem.

Creation

Deliverables

What will the project produce, build or deliver?

List the major deliverables in a table.

The deliverables for the project are listed in the table below, together with the estimated month of delivery.

Table of Deliverables

Provenance Tasks

Aa Name	# End month	≡ Description
Fellowship agreement signed	0	The fellowship agreement document, signed.
Partnership and AI report	2	A concise report outlining the findings from the first phase, after interactions with the stakeholders and potential partners. It will present what the needs and expectations of both groups are in connection with AI and Swarm storage.
SWIP proposal Toolkit specification	2	The initial specification of the Toolkit that is proposed in the form of a SWIP. The general features are: registering of data on Swarm (e.g. datasets and models), tracking all access and modifications, maintaining a record of provenance. It consists of: a set of smart contracts, services for automation of provenance, user interface for manual interaction.
Finalised Toolkit specification	3	Version of Toolkit specification that went through the SWIP process and contains relevant feedback and modifications.
Toolkit pre-release #1	6	First pre-release version of the Toolkit made according to the specification. Might exclude non-crucial features. Can be used in testing environments by interested parties (e.g. developers).
Marketing plan document	7	Document defining actions and tasks to promote the Toolkit and its adoption as well as gathering feedback.
Toolkit release #2	7	Second pre-release of the Toolkit that integrates crucial feedback from the first pre-release. Might exclude non-crucial features. Can be used in testing environments by interested parties (e.g. developers).
Beta toolkit launched	8	Beta release of the Toolkit that integrates crucial feedback from the second pre-release. Can be used in production environments.

The months of delivery provided in the table are estimates.

Disclaimer: The launched toolkit is considered Beta, as security audits are out of scope of the project and the Beta label reflects this fact.

Change

Plan

How and when will the work be carried out?

Include a plan where major task, phases, milestones are visible with the corresponding dates / timeline.

The deliverables together with their estimated delivery dates, relative to start of the project, are listed in the previous section.

Change

How are we going to engage stakeholders and manage the risks?

List major risks and their mitigation in a table. How can different stakeholders be best engaged?

Provenance Risks

Risk	Tags	Mitigation strategy
Funding is found not sufficient	Management	Explore options to secure additional funding during normal project path. Halt project is mismatch too great.
Development staggers due to technical difficulties	Development Requirements	Conduct thorough research before finalizing specifications through the SWIP process and decide on realistic, achievable goals.
Performance of the system is slow	Development	Optimize the system during the development phase, conduct performance testing, and use scalable infrastructure solutions to ensure high performance.
No adoption of tech by the community	Development Marketing	Engage with the community early through the SWIP process, gather feedback, and adjust the Platform to meet user needs. Conduct marketing and outreach to raise awareness and interest.
Regulatory risks due to decentralized storage	Management Requirements	Limit the project to supporting public data only, avoiding regulatory uncertainty. Stay informed about regulatory changes and adjust the project scope as needed.

Risk	Tags	Mitigation strategy
Development takes too long and funds run out	Development Management	Regularly review project progress and adjust scope if necessary to stay within budgetary constraints. Prioritize key features and functionalities to ensure completion.
Unclear scope and specs hit tech limits	Development Requirements	Clearly define the project scope during the SWIP process. Conduct feasibility studies to ensure that the technical requirements are achievable within project limits.
Low stakeholder engagement	Marketing	To effectively engage stakeholders and manage risks, a comprehensive communication plan will be established before release phase. This plan will outline the communication channels, methods for gathering feedback, and engagement strategies for each stakeholder group.