Metaverse Standards Forum
Licensing of Reputation Data for LLMs

Last Update: May 11, 2025

Status: Approved for Public Distribution

Version: 1.0

Reviewer	Due Date	Status	Contact
Digital Asset Management Working Group	December 17, 2024	Complete	digital_asset_management @lists.metaverse-standards.org
MSF Domains (Peer Review)	March 05, 2025	Complete	oversight@lists.metaverse-standards.org
Use Case Taskforce	May 11, 2025	Complete	use_case_task_force@lists.metaverse-standards.org

The purpose of this template is to provide a structured framework for collecting and documenting use cases within the Metaverse Standards Forum (MSF). Use cases are essential for understanding real-world scenarios where metaverse technologies are applied and where interoperability challenges may arise. This template guides MSF members in providing a concise yet comprehensive description of a use case, including its title, identifier, and summary. It also encourages contributors to list the benefits of the use case, identify actors or entities involved, and describe the use case scenario in detail, emphasizing interactions, challenges, and requirements. Additionally, it prompts the inclusion of relevant technical information, such as implementations, success metrics, and challenges faced. This template aims to facilitate the gathering of valuable use-case data to inform standards development and foster collaboration within the MSF community.

MSF members and MSF Domain Groups are invited to submit use cases.

NOTE: Organizations such SDOs who want to submit and add a use case would need a sponsor that is an MSF member. This process is established in order to have a contact person in MSF that can handle discussions and resolve open issues within regular meetings.

Eligible submitters:

- MSF Domain Groups
- MSF Members (Principal and Participant)
- External Organizations with Liaison Agreements (with the support of a MSF member that acts as sponsor)
- Standard Development Organizations (with the support of a MSF member that acts as sponsor)

Minimum Requirements for MSF Member Submissions not part of a Domain Group:

- Minimum required number of proposers: 3
- Minimum required number of supporters: 5

NOTE: Use cases submitted by SDOs and Liaison Organizations would also need to fulfill the same requirements (and would need a sponsor) unless they are submitted by a Domain Group.

MSF: *Metaverse Standards Forum*
POG: *Pre-qualified Organizations and Groups*
SPP: *Standards Related Publications and Projects*
DWG: *Domain Working Groups*
WG: *Working Group*
SDO: *Standards Development Organization*

Use Case Title

Licensing of Reputation Data for LLMs

Use Case Identifier

MSF2024-LLM-001

- Version 1.0
- Year of Release: 2025

Summary of Use Case

Description: This use case describes a framework for Licensing Personal and Reputation Data to organizations that develop Large Language Models (LLMs). The framework ensures controlled, transparent management of data usage, focusing on the protection of privacy, Intellectual Property Rights, and adherence to ethical standards in AI training. By facilitating clear data ownership and usage terms, this model aims to mitigate risks associated with data misuse and uphold data integrity within the AI sector.

Benefits:

- **Increased Transparency:** provides a transparent data usage framework that enhances trust and cooperation between Data Owners and LLM Developers.
- **Standards Compliance:** promotes the development and implementation of ethical guidelines in AI training, ensuring responsible use of data.
- **Robust Privacy Protections:** introduces stringent privacy safeguards that protect Personal and Reputation Data within LLM training environments.
- **Intellectual Property Management:** ensures that Data Owners retain control over their Intellectual Property Rights, safeguarding against unauthorized use.

- **Data Provider Compensation:** allows Data Owners to get compensated for contributing data to LLMs

Contributors and Supporters

- Digital Asset Management Working Group
- MSF Domains (Peer Review)
- Use Case Taskforce

Keywords

Data Licensing, Large Language Models, Intellectual Property Rights, Data Privacy, Ethical AI Training, Transparency in AI, Reputation Data Management

Actors/Entities

- **Data Owners**
  - **Human Users:**
    - **Avatar Data Owners:** individuals who create and manage their avatars, including appearance, behavioral patterns, and interaction histories within the Metaverse.
    - **Content Creators:** users who own original content such as virtual real estate, digital art, or In-Metaverse experiences.
  - **Commercial Entities:**
    - **Businesses and Brands:** Companies operating within the Metaverse, providing services or products, ranging from virtual goods to games to real-time interactive experiences. They manage data related to customer interactions, transactions, and behavioral analytics.
    - **Advertisers:** Firms that collect and analyze data on user behavior, preferences, and interactions to tailor marketing strategies within the Metaverse.
  - **Technology Providers:**
    - **Platform Developers:** entities responsible for developing and maintaining Metaverse Platforms. They handle vast amounts of data regarding user engagement, system performance, and user-generated content.
    - **Infrastructure Providers:** Companies that provide the necessary technology infrastructure, such as servers and networking solutions, crucial for the operation of Metaverse environments. They manage data related to network usage and operational telemetry.
- **LLM Developers:** Companies or research organizations that develop LLMs. They are responsible for acquiring Data Licenses from Data Owners, adhering to the terms of use, and implementing robust privacy measures to protect the data utilized in their models.

Metaverse STANDARDS FORUM™

Detailed Description of Use Case/Scenario

Preconditions:

- **Agreement Frameworks (optional):** actors have Licensing Agreements drafted, specifying terms of use, data handling, and compensation structures.

Main Flow:

1. **Data Licensing Request:** an LLM Developer identifies potential data sources within the Metaverse and sends a Data Licensing request to the Data Owners.
2. **Negotiation and Agreement:** if there are no Licensing Agreements already in place, Data Owners review the request and negotiate terms that respect their privacy, Intellectual Property, and compensation expectations.
3. **Data Preparation and Transfer:** Data Owners prepare the data as per the agreed standards and transfer it to the LLM Developer.
4. **Data Usage in LLM Training:** the LLM Developer uses the Licensed Data to train or refine its language models, ensuring all usage complies with the Licensing terms.

Alternative Flow

- **Rejection of Data Licensing Request:** if the Data Owners reject the Data Licensing request, the LLM Developer must either adjust the request to meet the Owners' conditions or seek alternative data sources.
- **Data Misuse:** if a breach of the Licensing Agreement is detected, such as unauthorized data sharing or usage beyond the scope, legal actions are initiated based on the predefined terms in the Licensing Agreement.

Postconditions

- **Performance Feedback and Adjustments:** Data Owners receive reports on the usage and performance of their data, allowing for adjustments in future Licensing Agreements.
- **Renewal or Termination of Agreement:** based on the success and compliance of the initial agreement, parties decide whether to renew, adjust, or terminate their Data Licensing Agreements.

Implementations and Demonstrations or Technical Feasibility

Existing Implementations

- **Dawex Data Exchange Technology:** Dawex provides a secure platform where organizations can license and exchange data. This platform is particularly effective in managing the rights, security, and traceability of data transactions, making it a suitable model for LLM Data Licensing.
- **Ocean Protocol's Blockchain-Based Data Sharing:** Ocean Protocol uses blockchain technology to facilitate safe and transparent Data Sharing and Licensing. Their approach ensures that data providers retain control over their data, with the ability to set conditions for its use, thereby providing a strong framework for ethical LLM training.
- **Pilot Project Between MIT and IBM:** A notable pilot project involved the Massachusetts Institute of Technology (MIT) and IBM, which focused on the ethical use of Personal Data

in AI research. This collaboration served to develop standards and practices that protect data integrity and user privacy during the AI model training process.

Technical Feasibility

- **Data Anonymization Techniques:** one specific technique widely recognized for ensuring data privacy is differential privacy, as applied by Google in their federated learning projects. Differential privacy provides a mathematical framework that quantifies how anonymized the data is, thus ensuring that individuals' data cannot be reverse-engineered or identified from large datasets.
- **Secure Data Transfer Protocols:** technologies like blockchain and encrypted data transmission methods provide robust solutions for secure data sharing between Data Owners and LLM Developers.

Challenges:

- **Data Standardization and Quality Assurance:** ensuring that the data provided by various owners is of high quality and adheres to standard formats necessary for LLM training can be challenging. Variability in data quality and format can impede the effectiveness of the models trained.
- **Legal Compliance:** complying with global and local privacy and data protection regulations such as General Data Protection Regulation (GDPR) in the EU and California Consumer Privacy Act (CCPA) in the US, and others poses a significant challenge, especially in the absence of operational guidelines that can help guide compliance and effective implementation.
- **Data Anonymization and Continuous Monitoring:** ensuring that data used in training does not violate privacy laws requires robust anonymization processes and continuous monitoring. Moreover, observing public (government) and private (corporate) mandates at a local, national, and global level, while considering time horizons (for e.g., short versus long term) when benchmarking and deciding the type of policy choices to adopt could be a great challenge.
- **Intellectual Property Rights Enforcement:** safeguarding the Intellectual Property Rights of Data Owners while promoting open collaboration and usage in the AI community is complex. Balancing these needs requires clear, enforceable Licensing Agreements and ongoing management.
- **Scalability of Licensing Processes:** as the demand for diverse datasets increases, scaling the Licensing processes while maintaining control and compliance becomes more difficult. Automated systems such as blockchain can help, but integrating these technologies presents its own challenges.
- **Ethical Use of Data:** ensuring that data is used ethically, particularly when it involves sensitive or Personal information, remains a persistent challenge. This involves not just legal compliance but also aligning with broader ethical standards set by the respective community and society.

Metaverse STANDARDS FORUM™

Requirements:

Technical and Functional Requirements

- **Advanced Anonymization Techniques:** deploy state-of-the-art data anonymization technologies to ensure compliance with privacy laws and protect individual identities.
- **Robust Security Measures:** integrate comprehensive cybersecurity measures to protect data from unauthorized access and breaches during data transfer and storage.
- **Automated Licensing Platforms:** develop or enhance platforms that can automate the Data Licensing process, reducing manual oversight and speeding up transactions.
- **Real-Time Compliance Monitoring:** establish systems that can monitor compliance with Licensing Agreements in real-time, utilizing technologies such as smart contracts on blockchain.
- **Feedback Mechanisms:** incorporate mechanisms for Data Owners to receive reports on the usage of their data and provide feedback on any issues or concerns.
- **Scalable Infrastructure:** design infrastructure that can handle increasing amounts of data and transactions without compromising performance or security.
- **Flexible Licensing Agreements:** create Licensing Agreements that are adaptable to different use cases and scalable as technology and regulations evolve.
- **Compliance with Regulatory Standards:** adhere to data protection regulations and ethical guidelines in all aspects of data handling and processing.

Interoperability Requirements:

- **API Integration:** ensure that Data Licensing Platforms and compliance monitoring systems can integrate seamlessly with existing enterprise and AI development environments via APIs.
- **Data Portability Standards:** adopt and promote standards that enable data portability between different platforms and systems, facilitating broader collaboration and use.
- **Data Standardization Tools:** implement tools and protocols for standardizing data formats and quality, ensuring uniformity across different data sources.

Other Key Considerations:

- **Privacy:** implement privacy enhancing technologies and practices to protect the privacy of Personal and Reputation Data during its collection, processing, usage, storage, and sharing – ensuring that access is restricted to authorized parties and solely for the intended LLM training purposes.
- **Cybersecurity:** robust cybersecurity measures to safeguard LLM Data from vulnerabilities, including unauthorized access, data breaches and illicit activities such as data trading or fraud.
- **Identity Verification:** reliable and secure Verification of LLM Developers' compliance assurances provided to Data Owners for Data Licensing qualification purposes, and to foster trust in the Digital Asset ecosystem.
- **Networking and Latency:** architect the system to minimize latency related to Data Licensing processes, ensuring smooth operation across geographically dispersed environments and diverse compliance requirements.

- **Ownership:** provide Data Owners with the ability to maintain oversight on their data usage, storage and sharing to ensure continuous compliance with Licensing Agreement.
- **Digital Ethics:** address ethical considerations by establishing or affiliating with an Ethics Board tasked with providing oversight, including regularly reviewing and guiding the ethical use of Licensed Data.
- **Provenance:** tools and protocols are needed to verify that the compliance assurances provided by LLM Developers to Data Owners – regarding terms of use, data handling, and compensation structures – are authentic and sufficient to qualify for Data Licensing.
- **Accessibility:** ensuring the Licensed Data is accessible to Data Owners from diverse backgrounds, with varying levels of technical expertise and accessibility requirements.

Relevant Domain Working Group (WGs):

- NA

Relevant Pre-qualified Organizations and Groups (POGs):

- NA

Relevant Specifications, Publications and Projects (SPPs):

- NA

Related Use Cases

- NA

Additional Comments

- This document is a living artifact and may be subject to revisions on a periodic basis to reflect the future state of Licensing of Reputation Data in the Metaverse, and or based on feedback received from MSF stakeholders that warrants an update in the future.