Enhancing Security, Transparency and Traceability in AI and Dataset Systems with SBOMs
Contents

1. Why AI+ Data sBOM Matter
2. Complexity increasing of AI System
3. AI+Dataset SPDX 3.0
4. AI Security (security, ethics, performance)
5. AI Licensing
6. AI Tools
7. What’s Next
8. Call to Action
AI Regional Regulations

- Estimated global annual cost of cybercrime of EUR 5.5 trillion by 2021.
- Huge opportunity actively evolving at fast pace

Measures for the Management of Generative Artificial Intelligence Services

“AI” and “governance” are slippery concepts.
Standard Regulations

- Estimated global annual cost of cybercrime of EUR 5.5 trillion by 2021.
- Over 300+ underway or published AI Standards

ISO/IEC 23894 – A new standard for risk management of AI

AI Standards: Roadmap for Ethical and Responsible Digital Environments
## AI + Data sBOM vs Standards

<table>
<thead>
<tr>
<th>AI-Data Fields 3.0</th>
<th>IEEE</th>
<th>ISO</th>
<th>EU Act</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>safetyRiskAssessment</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>sensitivePersonalInformation</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hyperparameter</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>energyConsumption</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>domain</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

Note: Detail chart: [https://docs.google.com/spreadsheets/d/1MnlUwrE29b8ODEz3H6Hl5HWAbK_N5vXy0_Pyry4HnOg/edit#gid=0](https://docs.google.com/spreadsheets/d/1MnlUwrE29b8ODEz3H6Hl5HWAbK_N5vXy0_Pyry4HnOg/edit#gid=0)
**EU Act Audit (LLMs)**

- Many issues exist even when comparing to one standard (EU AI Act) across most of the existing state-of-the-art AI models.
- In order to be able to establish compliance to several regulations around the world and identify potential risks associated with the use and adoption of AI software an automated standard, that captures all the required data and is machine readable is pivotal.
AI System Framework

Core
- Software
- Licensing
- Security
- Build

AI
- Model 1
- Model 2
- Model 3
- Model n

Data
- Data Training1
- Training Pile Data
- Prod Data
- Libraries
## AI/Dataset Profiles Definition

<table>
<thead>
<tr>
<th>Fact Sheets</th>
<th>None - Adhoc</th>
<th>Model Cards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oversimplification</td>
<td>IBM</td>
<td>Google</td>
</tr>
<tr>
<td>Data Flow / context</td>
<td>Deloitte</td>
<td>Hugging Face</td>
</tr>
<tr>
<td>Security and compliance</td>
<td>Lucidlink</td>
<td>OpenAI</td>
</tr>
<tr>
<td>Versioning update</td>
<td>Glasshouse</td>
<td>Meta</td>
</tr>
<tr>
<td>Environmental</td>
<td>CDW</td>
<td></td>
</tr>
<tr>
<td>Operational and maintenance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethical and Social implications</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SPDX AI Profile

Various details that we want to capture as a part of the profile

Properties

External property restrictions

Properties that are borrowed from other SPDX profiles

Used to express connections between other profiles and elements

Relationships

Required and optional fields

We propose a minimal and expansive list of properties to ensure easy adaptability
SPDX AI 3.0 (15 fields)

- energyConsumption: type: xsd:string
- standardCompliance: type: xsd:string
- Limitation: type: xsd:string
- typeOfModel: type: xsd:string
- informationAboutTraining: type: xsd:string
- informationAboutApplication: type: xsd:string
- Hyperparameter: type: /Core/DictionaryEntry
SPDX AI 3.0 (continued)

- modelDataPreprocessing: type: xsd:string
- modelExplainability: type: xsd:string
- sensitivePersonalInformation: type: /Core/PresenceType
- metricDecisionThreshold: type: /Core/DictionaryEntry
- Metric: type: /Core/DictionaryEntry
- Domain: type: xsd:string
- autonomyType: type: /Core/PresenceType
- safetyRiskAssessment: type: SafetyRiskAssessmentType
SPDX Data 3.0 (13 fields)

- datasetType: type: DatasetType
- dataCollectionProcess: type: xsd:string
- intendedUse: type: xsd:string
- datasetSize: type: xsd:nonNegativeInteger
- datasetNoise: type: xsd:string
- dataPreprocessing: type: xsd:string
SPDX Data 3.0 (continued)

- Sensor: type: /Core/DictionaryEntry
- knownBias: type: xsd:string
- sensitivePersonalInformation: type: /Core/PresenceType
- anonymizationMethodUsed: type: xsd:string
- confidentialityLevel: type: ConfidentialityLevelType
- datasetUpdateMechanism: type: xsd:string
- datasetAvailability: type: DatasetAvailabilityType
SPDX AI 3.0 (example LLAMA-2)

Core:  PackageName: LLAMA-2_Model
       PackageCreator: Meta
       PackageDownloadLocation: https://github.com/facebookresearch/llama
       PackageSupplier: Organization: Meta
       PackageVersion: 1.0

License:
       PackageConcludedLicense: LLMBA License
       PackageDeclaredLicense: LLMBA License

Build:
       BuiltDate: Noassertion
       ReleaseDate: September 2023
       ValidUntilDate: Noassertion
**SPDX AI 3.0 Example (Llama-2)**

- **Safety Risk Assessment:** Critical
- **Standard Compliance:** None
- **Limitation:** Carries risks with use. Testing conducted to date has been in English...
- **Type of Model:** LLMs - auto-regressive language model
- **Description:** Collection of pretrained & fine-tuned generative text models ranging from 7 to 70 billion tokens
- **Domain:** NLP
- **Training Info:** Trained between Jan -July 2023. This is a static model trained on an offline dataset.
- **Application Info:** Llama-2: Open Foundation and Fine-tuned Chat Models
- **Relationship Data used:** Pretrained on 2 trillion tokens of data but not available publicly.
- **Energy Consumption:** 3.3M GPU ...
- **Hyperparameters:** Unknown...
- **Metrics:** Benchmark - HumanEvalMBP, PIQA, SIQA, HellaSwag, WinoGrande, ARC ...
- **Metrics Decision thresholds:** Unknown
- **Preprocessing steps on data before model training:** Unknown
- **Model Update mechanism:** Download
- **Is Model Explainable:** No
- **Model Output:** Unknown
SPDX Data 3.0 Example (OASST1)

- **datasetType**: text
- **dataCollectionProcess**: Noassertion
- **intendedUse**: match Llama 2's prompt format as described [in this article](#).
- **datasetSize**: 1000 samples
- **datasetNoise**: Noassertion
- **dataPreprocessing**: dataset['train'].shuffle(seed=42).select(range(1000))
SPDX Data 3.0 Example (OASST1 -cont)

- **Sensor**: Noassertion
- **knownBias**: NoAssertion
- **sensitivePersonalInformation**: PII
- **anonymizationMethodUsed**: Noassertion
- **confidentialityLevel**: Low
- **datasetUpdateMechanism**: download
- **datasetAvailability**: public
Over 37,000 datasets metadata are captured using the SPDX dataset profile in order to conduct license compliance analysis as a part of the OpenDataology project.
AI Vulnerabilities Databases

- Security, ethics, and performance
- metadata, measurements, testcases benchmarks

AVID

Use of Copyright images without permission

DMCA 1202, which forbids the removal of copyright formation

OWASP 0.5

GitHub CoPilot lawsuit training and update of AI systems

AIAAIC

Tesla wins first Autopilot trial involving fatal crash

Napster was terrific, but it was also completely illegal,

EU AI Act “laws” as of April 2025
AI-specific security threats and a selection of controls

Development-time information security
- Data leak
- Data poisoning
- AI supply chain attack

Regular application security +
- Operations & governance
  - Operations
  - Governance
- Regular software engineering
  - Application development
  - Application code
- Model engineering
  - Train/test code
  - Trained regularities

Threats through use:
- Model theft
- Evasion
- Membership inference
- Model inversion
- Data disclosure
- Denial of model service

- Least model privilege
- Oversight
- Detect/filter abuse, Continuous validation
- Rate limit & monitor

Appsec threats:
- Standard attacks
- Model poisoning
- Insecure output
- Prompt injection
- Model theft
- Input leak

AI pipeline intellectual property leak (code, config)
- Parameters to perform attacks
- Model theft
- Model poisoning

// Source AI engineering framework: Software Improvement Group
1. **Data Poisoning Attacks:**
   - AI applications heavily rely on training data to learn and make decisions. Data poisoning attacks involve feeding malicious or manipulated data into the training dataset to influence the AI model's behavior.

2. **Adversarial Attacks:**
   - Adversarial attacks target AI models by introducing subtle changes to input data that can cause the model to misclassify or make incorrect decisions. These attacks are often challenging to detect and can have serious consequences, especially in critical AI applications like autonomous vehicles or security systems.

3. **Model Stealing:**
   - AI models, especially those deployed on cloud-based services, are susceptible to model stealing attacks. Attackers attempt to reverse-engineer and copy the AI model by sending queries to the service and observing the responses to build a replica of the model.

4. **Privacy Concerns:**
   - AI applications may process sensitive user data, leading to privacy issues if not handled properly. The model itself can inadvertently memorize sensitive information from the training data, leading to unintended data exposure.

5. **Overfitting and Generalization Issues:**
   - AI models can suffer from overfitting, where they perform well on the training data but poorly on new, unseen data. On the other hand, underfitting can also occur, resulting in poor performance on both training and new data.
1. **Explainability and Interpretability:**
   - Many AI models, especially deep learning models, are considered black boxes, meaning their decision-making process is not easily interpretable. Lack of explainability can raise ethical concerns and make it difficult to understand how and why the AI arrived at a particular decision.

2. **Transfer Learning Risks:**
   - Transfer learning, a technique where a pre-trained model is fine-tuned for a specific task, can lead to security risks. If the pre-trained model is compromised or contains biases, these issues can transfer to the application using the model.

3. **Backdoor Attacks:**
   - Attackers can insert backdoors into AI models during training, allowing them to trigger specific behaviors or manipulate the model's output in the future.

4. **Robustness Issues:**
   - AI applications may lack robustness when faced with input variations or novel scenarios that deviate from the training data distribution. This can lead to unexpected and potentially harmful outcomes.

5. **Dependency on Online Data and Services:**
   - Some AI applications rely on external online data or services, making them vulnerable to disruptions or changes in the availability of those resources.
Open Source Models & Datasets

Data: examples: Kaggle, GitHub, Hugging Face, Open Data Registry, GPT-*, Machine Learning Repository, Pile

Model examples: GPT-J(LLL) Bloom(LLL), Self Driving Car Simulation, Banking Loan, Product Recommendation, Image Caption Generator, Weather Forecasting, Hugging Face, Papers With Code
Other Data

- Pile contains 211,043,181 documents (unweighted),


- Syntecctic Data
AI Specific Licensing

Data Provenance Collection. 1,800 audited datasets suggests developers face legal and ethical risks. - 70% no data licenses, 15% incorrect.

<table>
<thead>
<tr>
<th>Artifact</th>
<th>Copyleft Permissive</th>
<th>Custom License</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data -</td>
<td>CC-BY-4.0, CLDA, Open Database License, Open Government License, <strong>ODC-PDDL, CC-0, ODC-BYCC-BY-SA</strong></td>
<td>Company specific open licenses</td>
</tr>
<tr>
<td>Model -</td>
<td>Open Access Model License, Apache 2.0l, MIT, i</td>
<td>Proprietary LLMs, Apache 2.0 with restrictions</td>
</tr>
</tbody>
</table>
AI Tools

- **Data License-generator** - Select license to use
- **Data Provenance Explorer** - Select data to use
- BiasAware - Assess Bias in Data
- Indiana University - PhD student - generate Sbom in excel & courseware on AI sBOM
What’s Next (Beyond 3.0)

- **Hardware** - AI specific requirements
- **Safety** - Verification and Validation
- **Working with OSS groups** to build AI + data sBOMs
- **New SPDX Fields**
  - LLM / Foundation Blocks
  - Empathy Data, Synthetic Data
  - Data Piles

Enhancements on 3.0
- integrated SBOM with AI and dataset
  - Several fields are still very textual,
  - Tooling support - Requires support from the tooling team to ensure tools are able to consume AI profile
- More user feedback
  - Open source project education for sbom
Call to Action

We want your feedback!

- What use cases do you need?
- What are we missing?

Ways to contribute:

- Join our meetings
  - SPDX AI & Data
  - Opendataloogy