SBOM Survey for ML Practitioners

Survey Flow

Standard: Consent (2 Questions)
Block: SBOM Background (5 Questions)
Standard: AIBOM Fields (7 Questions)
Standard: DataBOM Fields (10 Questions)
Standard: Challenges (7 Questions)
Standard: Demographic Questions (12 Questions)

Page Break

Start of Block: Consent

CF1 Investigating the state of practice in software supply chain management   RESEARCH GOAL AND PROCEDURE
The purpose of this study is to investigate issues, needs, and opportunities related to software supply chain management, specifically as it pertains **to machine/deep learning systems.**
  The (open source) software supply chain is the set of processes, components, and tools used to develop, build, and publish a software product (a system, library, tool, etc.). Software vendors often create software products by assembling open source software components (aka dependencies) developed by third-party developers or organizations. Common tasks involved in supply chain management include maintaining license compliance among components, managing and screening dependencies, and mitigating security threats introduced in software dependencies.
  If you decide to participate, you will take a brief survey/questionnaire. The survey should last about 15-20 minutes during which time you will be asked questions regarding your familiarity and experience with several topics related to the open source, machine learning supply chain, **including dependency management, Software Bills of Materials (SBOMs), security issues related to the supply chain, and open source licensing.**
  We may contact you by email to invite you to participate in a follow-up interview.
RISKS AND BENEFITS   There are no foreseeable risks for participating in this research. There are no benefits to you as a participant other than to further research in software engineering and the open source software supply chain.   PARTICIPATION   You must be at least 18 years old to participate.   Your participation is voluntary, and you may withdraw from the study at any time and for any reason. If you decide not to participate or if you withdraw from the study, there is no penalty or loss of benefits to which you are otherwise entitled. There are no costs to you or any other party. Your decision whether or not to participate will not prejudice your future relations with [Organization].   If you decide to participate, you will be entered into a drawing for a $50 USD Amazon gift card upon completion of this survey. We will randomly choose respondents and give out 10 gift cards.   CONFIDENTIALITY   The data collected by this study will be confidential, including your responses. Any information obtained in connection with this study that can be identified with you will remain confidential and disclosed only with your permission. You will be assigned a code number to protect your identity and all data will be kept secure. If you give us your permission by signing this document, we plan to disclose the results of the questionnaire in any publication resulting from this study. The disclosed results will not be personally identifiable (if needed, they will be anonymized). The de-identified data could be used for future research without additional consent from participants.   The Institutional Review Board (IRB) and [committee] that monitor research on human subjects may inspect study records during internal auditing procedures and are required to keep all information confidential.   CONTACT   This research is being conducted by [Anonymous Author(s)] from [Organization].
Questions regarding the rights of research subjects may be directed to [redacted].
Such committee has reviewed and approved the present research

([protocol]) CONSENT You are welcome to print this page to keep a copy of this form.
Do you consent to participate in this survey?

JS

CF2 **YOU ARE MAKING A DECISION WHETHER OR NOT TO PARTICIPATE.**

IF YOU WANT TO PARTICIPATE, PLEASE ENTER YOUR NAME IN THE TEXT FIELD BELOW, AND START THE SURVEY.

Please enter your name.

____

End of Block: Consent

Start of Block: SBOM Background

B1 Are you familiar with the concept of Software Bill of Materials (SBOM)?

○ Yes (1)

○ No (2)

Page Break

Definition 1 The NTIA defines a **Software Bill of Materials** (SBOM) as follows:

An SBOM is a formal, machine-readable **inventory of software components and dependencies**, information about those components, and their hierarchical relationships. These inventories should be comprehensive – or should explicitly state where they could not be.

SBOMs may include **open source or proprietary software** and can be widely available or access-restricted. SBOMs should also include baseline attributes with the ability to uniquely identify individual components in a standard data format. The most efficient generation of SBOMs is as a byproduct of a modern development process. For older software, less-automated methods exist.

Definition 2 In this survey, an **AIBOM** refers to a Software Bill of Materials (SBOM) created for an Artificial Intelligence (AI) system, specifically, **a machine/deep learning (ML/DL) system.**

Display This Question:

If Are you familiar with the concept of Software Bill of Materials (SBOM)? = Yes

B2 Which of the following SBOM formats are you familiar with? Select all that apply.

□ SPDX (1)

□ CycloneDX (2)

□ SWID (3)

□ Other (please specify) (4) ____

□ ⊗None of the above (5)

Display This Question:

If Are you familiar with the concept of Software Bill of Materials (SBOM)? = Yes

B3 Are you aware of any existing SBOM standards / specifications for ML/DL systems (AIBOMs)?

○ Yes (please list them) (1) ____

○ No (2)

End of Block: SBOM Background

Start of Block: AIBOM Fields

Display This Question:

If Are you familiar with the concept of Software Bill of Materials (SBOM)? = Yes

And Which of the following SBOM formats are you familiar with? Select all that apply. != None of the above

AI1 Indicate your agreement with the following statement: "Current SBOM formats adequately support machine/deep learning (ML/DL) systems."

○ Strongly disagree (1)

○ Disagree (2)

○ Neutral (3)

○ Agree (4)

○ Strongly agree (5)

AI2 What data fields should be included in SBOMs to adequately describe a machine/deep learning (ML/DL) system? Select all that apply.

□ Environment in which the model was trained in (1)

□ Model structure / architecture (2)

□ Model hyperparameters (3)

□ Model parameters (4)

□ Known model/data defects (5)

□ Potential model biases (6)

□ Potential fairness issues (7)

□ Unique model identifier (8)

□ Optimizers, loss functions, etc. (9)

□ Description of the training data (10)

□ Description of the validation and testing data (11)

□ Runtime performance requirements (12)

□ Data (pre-)processing techniques (13)

□ Presence of Personally Identifiable Information (14)

□ Model version (15)

□ Model description (16)

□ Creation / building time (17)

□ Model license (18)

□ Data version (19)

□ System dependencies (20)

□ Others (please specify) (21) ____

Display This Question:

If Are you familiar with the concept of Software Bill of Materials (SBOM)? = Yes

And Which of the following SBOM formats are you familiar with? Select all that apply. != None of the above

AI3 Can SBOMs of ML/DL systems (AIBOMs) follow the same current SBOM specifications and standards?

○ Yes, current SBOMs can be adapted to support AI systems (1)

○ No, an entirely new solution is necessary (2)

○ I don't know (3)

Display This Question:

If Can SBOMs of ML/DL systems (AIBOMs) follow the same current SBOM specifications and standards? = No, an entirely new solution is necessary

AI4 Please describe a potential new solution for specifying SBOMs of ML/DL systems (AIBOMS).

____

____

____

____

____

AI5 Are you aware of any tools which currently facilitate the automatic generation of AIBOMs?

○ Yes (please list them) (1) ____

○ No (2)

AI6 How would you prefer AIBOMs be distributed? Rank by preference (by dragging the options).

_______ Stored with trusted third party (1)
_______ Located in repositories along with source code or model (2)
_______ Downloadable from project website (3)
_______ Available from the project developers upon request (4)
_______ Attached to software binaries (5)
_______ Other(s) (6)

AI7 How can we ensure that AIBOMs completely and correctly report all the dependencies of ML/DL systems?

____

____

____

____

____

End of Block: AIBOM Fields

Start of Block: DataBOM Fields

Definition 3 In this survey, a **DataBOM** refers to a Bill of Materials (BOM) created for a dataset used to train, validate, and/or test **a machine/deep learning (ML/DL) system.**

D1 Indicate your agreement with the following statement: "Datasets for ML/DL systems should come with a bill of materials (**DataBOM**)."

○ Strongly disagree (1)

○ Disagree (2)

○ Neutral (3)

○ Agree (4)

○ Strongly agree (5)

Display This Question:

If Are you familiar with the concept of Software Bill of Materials (SBOM)? = Yes

D2 Are you aware of any existing standards for DataBOMs?

○ Yes (please list them) (1) ____

○ No (2)

D3 What data fields should be included in a DataBOM to adequately describe a dataset for a ML/DL system? Select all that apply.

□ Data sources (1)

□ Data pre-processing steps (2)

□ Data transformations (3)

□ Dataset authors / creators (4)

□ Data labelers / taggers (5)

□ Procedure for data collection and curation (6)

□ Presence of Personally Identifiable Information (7)

□ Dataset statistics (8)

□ Dataset version (9)

□ Dataset license (10)

□ Dataset description (11)

□ Version description (dataset change log) (12)

□ Creation time (13)

□ Known data biases (14)

□ Known vulnerabilities and issues (15)

□ Unique identifier (16)

□ Types of data included in the dataset (image, text, binary, etc) (17)

□ Dataset size (18)

□ Others (please specify) (19) ____

D4 Are you aware of any tools that generate DataBOMs?

○ Yes (please list them) (1) ____

○ No (2)

D5 What do you think the relationship between AIBOMs and DataBOMs should be?

____

____

____

____

____

D6 How should DataBOMs be shared / distributed? Rank by preference (by dragging the options).
_______ Stored with a trusted third party (1)
_______ Located alongside the dataset (2)
_______ Downloadable from maintainer or owner website (3)
_______ Available from owner upon request (4)
_______ Other(s) (5)

D7 What are some of the benefits you expect to see from using DataBOMs, if any?

____

____

____

____

____

D8 What main challenges do you foresee in the creation and use of DataBOMs?

____

____

____

____

____

D9 How can we ensure that DataBOMs completely and correctly report all the dependencies and/or components of datasets used by ML/DL systems?

____

____

____

____

____

End of Block: DataBOM Fields

Start of Block: Challenges

C1 What are some of the benefits you expect to see from using AIBOMs, if any?

____

____

____

____

____

C2 What main challenges do you foresee in the creation and use of AIBOMs?

____

____

____

____

____

C3 Do vulnerability databases (such as CVE) exist for ML/DL systems?

○ Yes (please list them) (1) ____

○ No (2)

○ I don't know (3)

Page Break

Display This Question:

If Do vulnerability databases (such as CVE) exist for ML/DL systems? = Yes (please list them)

C4 Indicate your agreement with the following statement: "Vulnerability databases for ML/DL systems meet the security needs of developers and consumers"

○ Strongly disagree (1)

○ Disagree (2)

○ Neutral (3)

○ Agree (4)

○ Strongly agree (5)

C5 Do vulnerability databases (such as CVE) exist for datasets of ML/DL systems?

○ Yes (please list them) (1) ____

○ No (2)

○ I don't know (3)

Display This Question:

If Do vulnerability databases (such as CVE) exist for datasets of ML/DL systems? = Yes (please list them)

C6 Indicate your agreement with the following statement: "Vulnerability databases for datasets of ML/DL systems meet the security needs of developers and consumers"

○ Strongly disagree (1)

○ Disagree (2)

○ Neutral (3)

○ Agree (4)

○ Strongly agree (5)

C7 Given recent concerns regarding scraped public material being included in machine/deep learning datasets, how should AIBOMs address licensed material, if at all?

____

____

____

____

____

End of Block: Challenges

Start of Block: Demographic Questions

*

Q1 How many years of experience developing machine/deep learning systems do you have?

____

Q2 How would you describe your primary role?

○ Programmer (1)

○ Tester (2)

○ Project Lead (3)

○ Project Manager (4)

○ IT Manager (5)

○ DevOps Engineer (6)

○ MLOps Engineer (7)

○ ML/DL Engineer (8)

○ Data scientist (9)

○ Other (please specify) (10)

____

Q3 What is your highest level of education?

○ Did not graduate from high school (1)

○ High School (2)

○ Some college (3)

○ Bachelor's Degree (4)

○ Master's Degree (5)

○ Doctoral Degree (6)

○ Other (please specify) (7)

____

Q4 Which programming languages have you most used in past projects? Select all that apply.

□ Python (1)

□ Java (2)

□ C / C++ (3)

□ Javascript (4)

□ C# (5)

□ Others (please specify) (6) ____

Q5 Which machine / deep learning models do you use most frequently?

○ Deep Learning models (CNNs, RNNs, Transformers, etc.) (1)

○ Non-deep learning models (SVMs, Decision Trees, etc.) (2)

○ Both deep and non-deep learning models (3)

○ Other (please specify) (4) ____

Q6 Do you primarily work on open source or closed source projects?

○ Open source only (1)

○ Closed source only (2)

○ Open and closed source (3)

Q7 For which domains (e.g., banking or healthcare) have you developed applications / systems?

____

Q8 Do you have a background in computer security? Formal (college classes, degree, certification) or informal (self-learning or other training)?

○ Yes, formal (1)

○ Yes, informal (2)

○ No (3)

Q9 Do you have a background in software licensing? Formal (college classes, degree, certification) or informal (self-learning or other training)?

○ Yes, formal (1)

○ Yes, informal (2)

○ No (3)

Q10 How often did you release or help release a new major version of an application / software over the past two years?

Please give your best estimate; if you develop more than one software / application, please answer based on the most frequently updated software / application.

○ Never (1)

○ Annually (2)

○ Quarterly (3)

○ Monthly (4)

○ More frequently (5)

Q11 Which country / countries are you and / or your organization based in?

____

*

Q12 Please enter your email address. This will be used to contact you for compensation. If we are interested in your responses, we may contact you for a follow-up interview, if you are willing.

____

End of Block: Demographic Questions