Data Integrity Threat Model

Ishan Abhinit (iabhinit@iu.edu)

Von Welch (vwelch@iu.edu)

19 Aug, 2022 (version 1.1.4)

About this document

This document uses the OSCRP[1] as a point of reference and for guidance.

Guaranteeing the integrity of scientific workflow processing on distributed cyberinfrastructure that is prone to data corruption and malicious attacks is of paramount importance. Data driven application depends on the integrity of underlying scientific computational workflow and on the integrity of associated data products.

The Scientific Workflow Integrity with Pegasus (SWIP) project[2] is working to improve the security and integrity of scientific data by integrating cryptographic integrity checking and provenance information into the Pegasus workflow management system (WMS)[3].

The SWIP project addressed the integrity checking to make sure that scientific workflow computations are free from integrity errors. SWIP project does not address the analysis of integrity errors found in the process i.e. tracing the source of the error or doing the root cause analysis to remedy the underlying cause. That is the goal of **IRIS** project[4] i.e. to detect, diagnose, and pinpoint the source of unintentional integrity errors in the scientific workflow executions on distributed cyberinfrastructure.

This document applies the Open Science Cyber Risk Profile (OSCRP) to determine the threats that are in scope for the IRIS project. It is assumed the reader of this document is familiar with scientific workflows – if not, the Pegasus Overview webpage[5] is a good starting point. A brief overview of the OSCRP process is given in this document, with a reference to the complete profile.

Acknowledgements

The IRIS Project is supported by the National Science Foundation under Grants 1839900. SWIP is funded by NSF award 1642070.

[1] https://trustedci.org/oscrp/
[2] https://cacr.iu.edu/projects/swip/
[3] https://pegasus.isi.edu/
[4] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1839900&HistoricalAwards=false
[5] https://pegasus.isi.edu/about/

The Open Science Cyber Risk Profile (OSCRP) is a joint effort of Trusted CI and the Department of Energy's Energy Sciences Network (ESnet). Trusted CI, the NSF Cybersecurity Center of Excellence is supported by the National Science Foundation under Grant ACI-1547272.

The views expressed in this document do not necessarily reflect the views of the National Science Foundation or any other organization.

Scope and Assumptions

Scope: The goals of the IRIS project are to:

> Provide assurance that sources of integrity errors can be detected and remedied.

The OSCRP Process

In this section, a brief overview of the OSCRP process is given. This document only undertakes the first four steps of the profile, up to the generation of concerns. The complete OSCRP profile is available at:

> Peisert, Sean, Von Welch, Andrew Adams, RuthAnne Bevier, Michael Dopheide, Rich LeDuc, Pascal Meunier, Steve Schwab, and Karen Stocks. 2017. Open Science Cyber Risk Profile (OSCRP), Version 1.2. March 2017. http://hdl.handle.net/2022/21259

OSCRP process, in brief:

1. **Identify the stakeholders** of the science project — at the very least, this includes the principal investigator(s) and science team; other researchers, including possible external users; the institution that owns or manages the science instrument, the project, and the mission it supports; and possibly human subjects of the science project.
2. **Create an Asset inventory** for the project by looking through the list in the "Common Open Science Assets" section, and identifying all the Assets relevant to the Open Science project.
3. For each mission critical science Asset, examine the Concerns, Consequences, and Avenues of Attack diagram associated with the Asset and note which **Concerns and Consequences** are relevant to the project, and the extent to which they are relevant
4. For each relevant Concern note the **Vectors of Attack** that could cause the Concern to be realized.
5. Agree on and implement controls to mitigate Avenues of Attack. This step is out of scope of this document.
6. Repeat this assessment annually. This step is out of scope of this document.

Each of the first four steps of the OSCRP Analysis is captured in the following sections.

OSCRP Step 1: Identify Stakeholders

Stakeholders are those who are concerned about the security and integrity of the scientific workflow.

1. Researcher: This is the workflow initiator who defines the workflow. They are presumed to be part of a research team who will leverage the results of the workflow in their science. They will use the workflow results to support scientific outcomes.

2. Data Consumers: These are researchers or others who may consume the data produced by a workflow either in subsequent workflows as inputs or to validate the results of the Researcher.

3. Owner of the IRIS project: The institution that owns and manages the IRIS project.

OSCRP Step 2: Asset Inventory

Assets are IT entities (computers, data, etc.) whose integrity is necessary to SWIP's goals. Section 8 of the OSCRP has a list of common assets which we map to.

1. **Transient Workflow Data**: Transient data products created during the course of the workflow execution and that do not persist meaningfully past the completion of the workflow. E.g. intermediate data products generated by one job in the workflow and consumed by the subsequent job and then discarded.
    - We treat Transient Workflow Data to OSCRP Internal Data.
2. **Data Products**: Data that is taken as input by a workflow or generated by a workflow and expected to persist and be consumed by Data Consumers.
    - We treat Data Products as OSCRP Public Data (To consider confidentiality, one could treat this as OSCRP Embargoed Data).
3. **Metadata**: Data created by a workflow execution describing the workflow execution and resulting data products. This includes integrity information (e.g. hashes) about the Data Products added by SWIP project enhancements. Metadata is expected to persist in perpetuity.
    - We treat Metadata as OSCRP Accounting Information.
4. **Researcher System**: The interface used by the Researcher to craft and initiate the workflow. Specifically, it executes Workflow Managed System client, holds abstract workflow description and workflow plan, holds metadata (including integrity information) regarding workflows, and holds Researcher credentials for accessing Computational and Data Storage Systems.
    - We treat the Researcher system as an OSCRP Desktop.
5. **Workflow Management System (WMS):** This system translates abstract workflow description from Researcher into work plan, maps workflow plan to Computational and Data Storage Systems, manages workflow execution, and orchestrates the creation and checking of data integrity information.
    - We treat the WMS as an OSCRP Workflow.

6. **Computational Systems:** Computer systems that run computational aspects of the workflow, provide for temporary data storage during workflow (during computation and stage-in/out), provide the software stacks for workflow execution, and create and check data integrity information within the workflow.
    - We treat Computational Systems as OSCRP Servers.
7. **Data Storage Systems:** Computer systems that provide for long-term storage of data consumed by and produced by workflows, and serve to make data available to Data Consumers.
    - We treat Data Storage Systems as OSCRP File Stores.
8. **Network System:** The IT system that transports data between the Research Systems, Workflow Management System, Computational Systems, and Data Storage Systems involved in executing the workflow.
    - We treat the Network System as an OSCRP Network.

OSCRP Step 3: Concerns and Avenues of Attack

Table below details the analysis of Assets identified in the previous section, using the OSCRP, to arrive at the following set of concerns. For each concern, we discuss the relation to IRIS scope of workflow and data integrity.

Concerns:

1. Transient Workflow Data, Data Products, or Metadata being corrupted or lost by issues with data processing
    - Locating the source of such data corruption is the goal of IRIS
    - A lack of availability of data is not in IRIS's scope.
    - A lack of availability of metadata means that corruption of data may go undetected and hence is a concern to IRIS.

2. Data Products being falsely created by malicious insiders (Not in scope).
    - Protecting against the creation of false data by malicious insiders is ~~not~~ a goal of IRIS since we are looking only at non-malicious/unintentional integrity failures.

3. Data Products being falsely created due to misconfiguration.
    - Data being falsely created due to misconfiguration is in the scope of the IRIS project.
    - False data would lead to incorrect output for scientific workflows.

4. Researcher System being unavailable due to loss, unintentional damage, network unavailability, or misconfiguration.
    - A lack of availability of metadata means that corruption of data may go undetected and hence is a concern to IRIS as the metadata is stored on the researcher's system.

5. Workflow Management System running flawed processes due to human interference, storage, communications or other data issues.
    - Locating the source of the integrity error is a goal of IRIS.

6. Computational Systems inaccessible or not performing as expected due to misconfiguration or network issues (communication issues).
    - Availability of computational systems is not a concern of IRIS.
    - Locating the source of errors due to computational systems not performing as expected and producing corrupted results is a goal of IRIS.

7. Data Storage Systems being unavailable due to theft or intentional damage.
    - Availability of data storage systems, and any Data Products they are storing, is not a concern of IRIS.

8. Data Storage Systems not performing as expected due to damage, or misconfiguration.
    - Locating the source of Data Storage Systems not performing as expected and corrupted Data Products is a goal of IRIS.

9. Network Systems not transporting data due to unintentional damage, or misconfiguration.
    - Network Systems failing to transport data is in scope of IRIS.

10. Network Systems altering data during transport due to damage or tampering.
    - Locating the source of altered data being transported is a goal of IRIS.

Table 1: Detailed Analysis of Assets to Concerns and Avenues of Attack

Assets	Concern(s)	Consequence/Impact	Avenue of Attacks
Transient Workflow (Internal Data)	Corrupted data, incorrect data or lost data	Scientific Workflow producing incorrect/invalid results	Issues with sensor equipment, issues with data processing
Data Products (Public Data)	Falsely created data due to misconfiguration	Scientific Workflow producing incorrect/invalid results	Issues with sensor equipment, issues with data processing
Metadata (Accounting Data)	Corrupted data, incorrect data or lost data	Scientific Workflow producing incorrect/invalid results	Issues with sensor equipment, issues with data processing
Researcher System(Desktop)	Loss of Data Products and Metadata stored on the Researcher Systems	Scientific Workflow cannot be initiated or executed	Unintentional loss or damage or unreachable
Researcher System(Desktop)	Device not performing as expected	Scientific Workflow producing incorrect/invalid results	Misconfigured or damaged equipment
Workflow Management System (Workflow)	Lost or incorrect process	Scientific Workflow producing incorrect/invalid results	Issues with storage
Computational Systems (Servers)	Device Inaccessible	Scientific Workflow cannot be executed	Unintentional loss or damage or unreachable
Computational Systems (Servers)	Device not performing as expected	Scientific Workflow producing incorrect/invalid results	Misconfigured, lost or damaged equipment
Data Storage Systems (File Stores)	Service unreachable or service unavailable	Archived workflow inaccessible	Service Provider’s Storage systems files getting corrupted or lost
Data Storage Systems (File Stores)	Service not performing as expected	Scientific Workflow cannot be initiated/executed	Service Provider’s Storage systems files getting corrupted or lost
Data Storage Systems (File Stores)	Service not performing as expected	Erroneous result/Invalidated result/	Broken communication

		Scientific Workflow cannot be initiated or executed	between servers and file stores
Network Systems (Network)	Data can't be transported	Scientific Workflow cannot be initiated/executed	Network equipment lost or damaged
Network Systems (Network)	Data inconsistency	Scientific Workflow cannot be initiated/executed	Wireless interference, microwaves, atmospheric conditions

OSCRP Step 4: Vectors of Attack/Failure

Table 2: Vectors that could cause the concern to be realized

Concerns	**Vectors of Attack/Failure**
Corrupted data, incorrect data or lost data (Internal data)	Bit Rot (due to wear, dust, temperature, background radiation)
Falsely created data due to misconfiguration	Misconfiguration
Loss of Data Products and Metadata stored on the Researcher System	Bit Rot (due to wear, dust, temperature, background radiation), software bugs, natural disasters, blackouts
Device(Researcher System) not performing as expected	Physical hardware issues, Bit Rot (due to wear, dust, temperature, background radiation), software bugs
Device Inaccessible (Servers)	Physical hardware issues, Bit Rot (due to wear, dust, temperature, background radiation), software bugs
Device(Servers) not performing as expected	Physical hardware issues, Bit Rot (due to wear, dust, temperature, background radiation), software bugs
Service unreachable or service unavailable (File stores)	Bit Rot (due to wear, dust, temperature, background radiation), network unavailable, Failed transfers, software bugs, communication timeout
Data can't be transported(Network)	Physical hardware issues, Failed transfers, software bugs

Data Inconsistency	Wireless interference, microwaves, atmospheric conditions
Physical Integrity of data compromised	Natural disaster, blackouts

OSCRP Step 5: Mitigations

This is not in the scope of this document.

The goal of the SWIP project was to provide assurances that a scientific workflow is not accidentally or maliciously tampered with during its execution. To achieve this goal, SWIP integrated data integrity protection into the Pegasus WMS. Doing this, allowed us to take advantage of two aspects of WMSes:

1. Routine tasks can be automatically handled such as generating and verifying integrity data that human researchers find tedious and error prone.
2. They have a holistic view of workflow allowing for integrity verification from end-to-end, also catching errors that occur between storage and transfer technologies. These aspects of WMS are leveraged to provide data integrity via cryptographic hashes.

We added new capabilities to Pegasus WMS to automatically generate and track checksums for both when input files are introduced and the files that are generated during execution. The first production release with integrity protection enabled by default was released with pegasus 4.9.0 on October 31st, 2018.

Additionally, SWIP project also developed Chaos Jungle - a toolkit that provides a controlled environment for integrity experiments by allowing researchers to introduce a variety of predictable network errors. This was done to make sure that our integrity protection work. Chaos jungle allows us to predictably introduce errors, ensuring that our integrity protection is working, and assess how this protection mechanism impacts performance.[6]

[6] Mats Rynge, Karan Vahi, Ewa, Deelman, Omkar Bhide, Randy Heiland, Von Welch, Raquel Hill, Anirban Mandal, Ilya Baldin, William L. Poehlman and F. Alex Feltus. 2018. *Integrity Protection for Scientific Workflow Data: Motivation and Initial Experiences*

Asset Mapping Diagram:

Figure 1 illustrates Pegasus Workflow Management System (WMS) which provides a comprehensive model for representing the workflow of a scientific application and indicates where OSCRP Asset integrates into the Pegasus WMS architecture. A key goal of SWIP is to provide assurance that input and output data associated with a given workflow can be detected. And as mentioned above, in IRIS, we are trying to pinpoint the source of these errors with respect to the scientific workflow architecture.

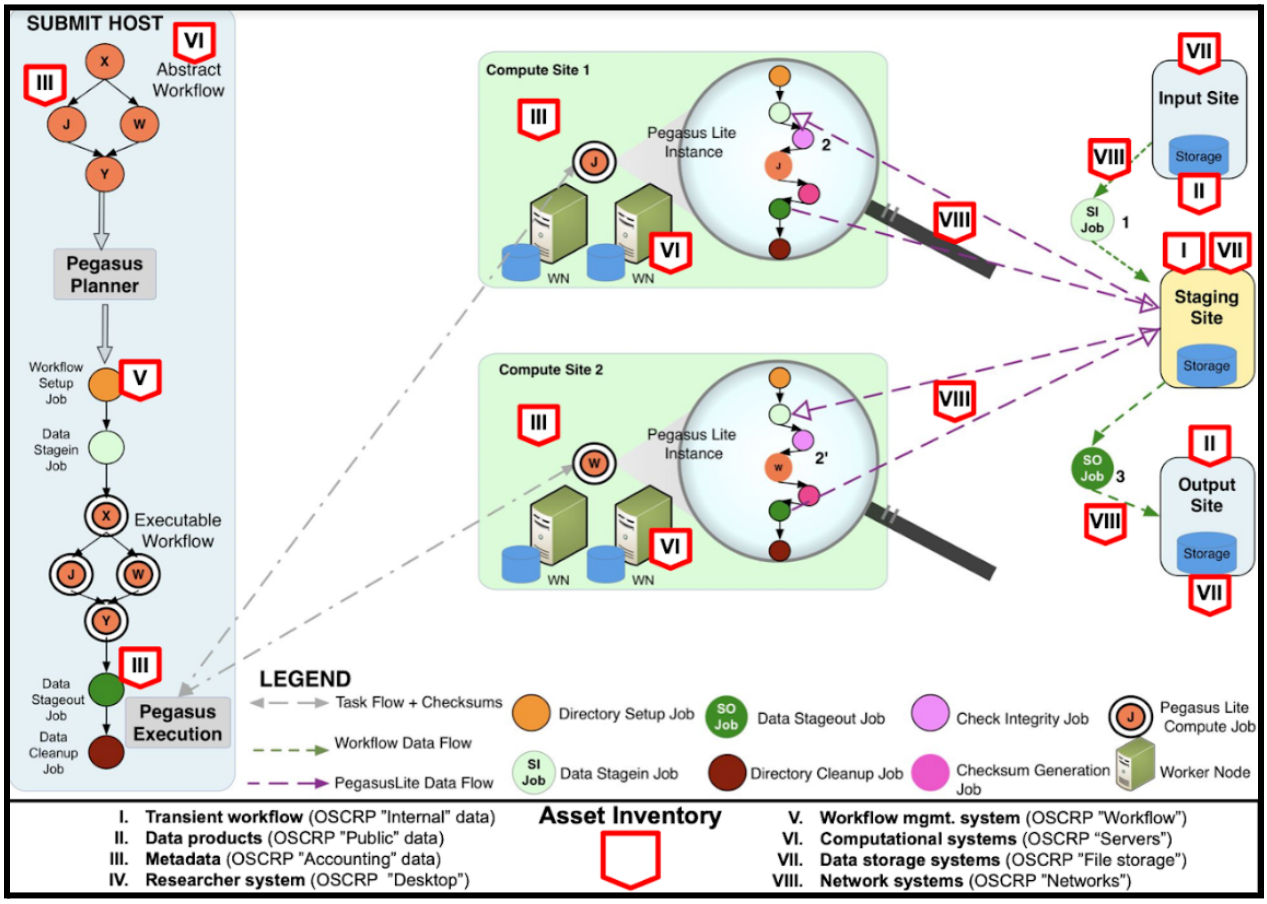

Figure 1. Pegasus Workflow Management System (WMS)