# Trustworthy repository for RHIC DAP

RHIC DAP Round Table

01/22/2024

# Accurately preserving knowledge

- Future users of RHIC data may (will) lack the context or memory of the experiment.
- Current internal websites cannot be presented as such to future users
  - Contain outdated, inaccurate (and sensitive?) information accumulated over time
  - Implemented with software needing long-term support.
- Cleaning unpublished documentation before preserving is essential
  - Only the experiments can perform this cleaning.
  - It is also an opportunity to review it and identify any missing information.
  - *Include reviewing documentation as part of the process for publishing analyses?*
  - An opportunity to migrate cleaned content to a static, simplified website.
- The goal is to establish a Trustworthy Repository for RHIC data

# What is a Trustworthy Repository?

- It is a digital archive that meets established standards and practices to ensure the long-term preservation, accessibility, and reliability of its digital assets.
  - Digital assets: data, code, notes, figures, documentation, papers,…
- Trustworthy repositories are crucial for safeguarding digital data, ensuring it remains usable and reliable over time, even as technology evolves.
- Trustworthy repositories are critical components of modern digital infrastructure, supporting the preservation and accessibility of knowledge for generations.
- Trustworthy repositories are important for:
  - Scientific Integrity: Preserves data for verification, reproducibility, and future research.
  - Cultural Heritage: Protects digital assets of cultural and historical significance.
  - Regulatory Compliance: Ensures data preservation mandates and intellectual property compliance.

# Key Characteristics of a Trustworthy Repository

- **Authenticity**: Maintains the authenticity of digital objects, ensuring they are preserved without unauthorized alterations.
- **Integrity**: Ensures stored data's integrity through checksums and version control.
- **Accessibility**: Guarantees access to data for designated users or communities over time, regardless of changes in technology.
- **Sustainability**: Has a clear mission, adequate resources, and a long-term strategy to support preservation efforts.
- **Transparency**: Maintains clear documentation of policies, procedures, and technical implementations.
- **Compliance with Standards**: Adheres to internationally recognized frameworks like the OAIS model (ISO 14721) and standards such as ISO 16363 or CoreTrustSeal
- **Security**: Implements robust security measures to protect against data loss, unauthorized access, and cyber threats.
- **Community Engagement**: Engages with its user community to ensure the repository meets their needs and expectations.

# CoreTrustSeal & The OAIS Model

- CoreTrustSeal is a certification standard for trustworthy digital repositories. It ensures that they meet specific criteria for the long-term preservation of digital objects.
- The OAIS model is a framework defined by ISO 14721:2012 that outlines the responsibilities and functions of an archival system to preserve digital information over the long term.
- CoreTrustSeal certification ensures a repository adheres to the OAIS model's standards, promoting the best digital preservation and data management practices.
- Over 160 certified repositories from many science fields

# The OAIS Model Functional Entities

**Ingest**: Accepts SIPs and prepares them for archiving.

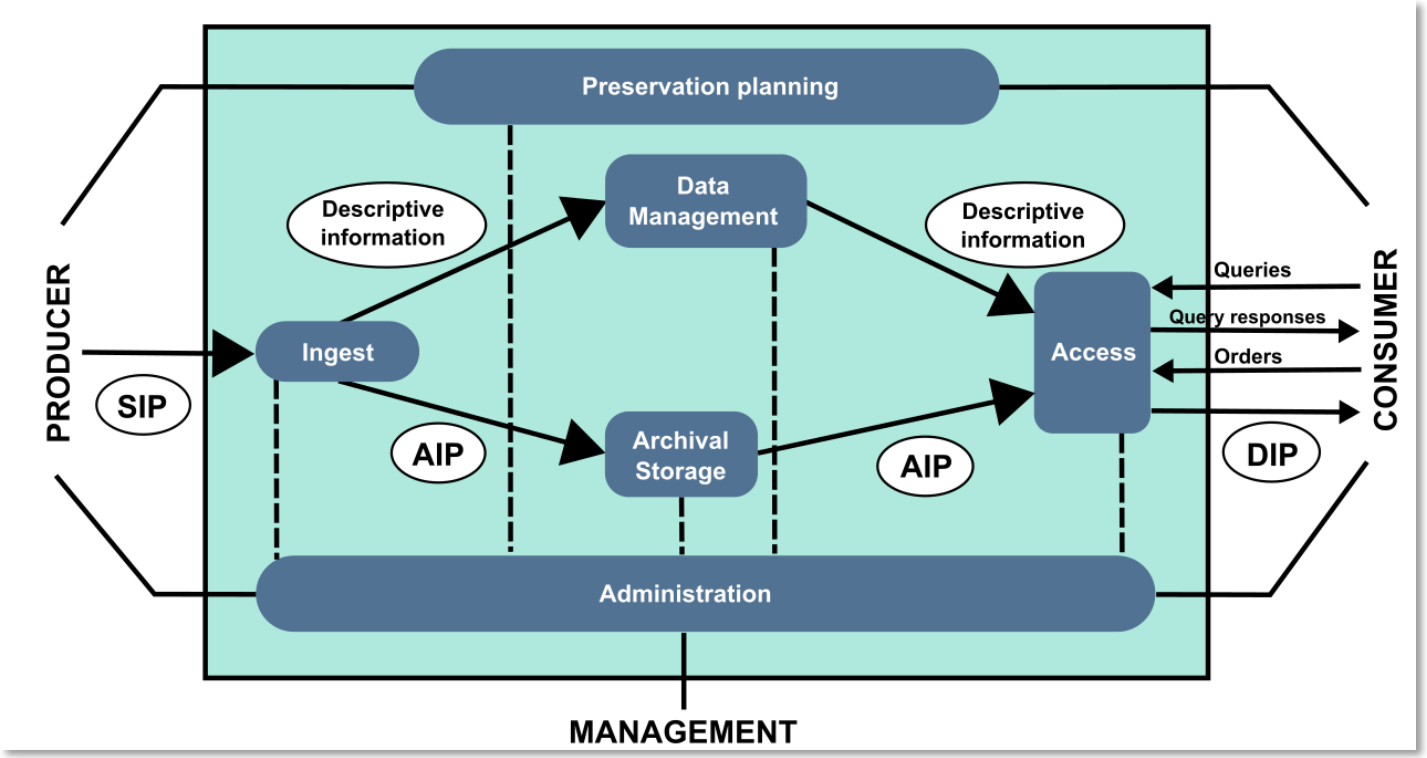
**Archival Storage**: Stores and preserves AIPs, ensuring their integrity and usability over time.

**Data Management**: Manages metadata, indexes, and retrieval mechanisms.

**Access**: Facilitates user queries and delivers DIPs.

**Preservation Planning**: Ensures content remains accessible by addressing technological obsolescence.

**Administration**: Manages the overall operation of the archive.



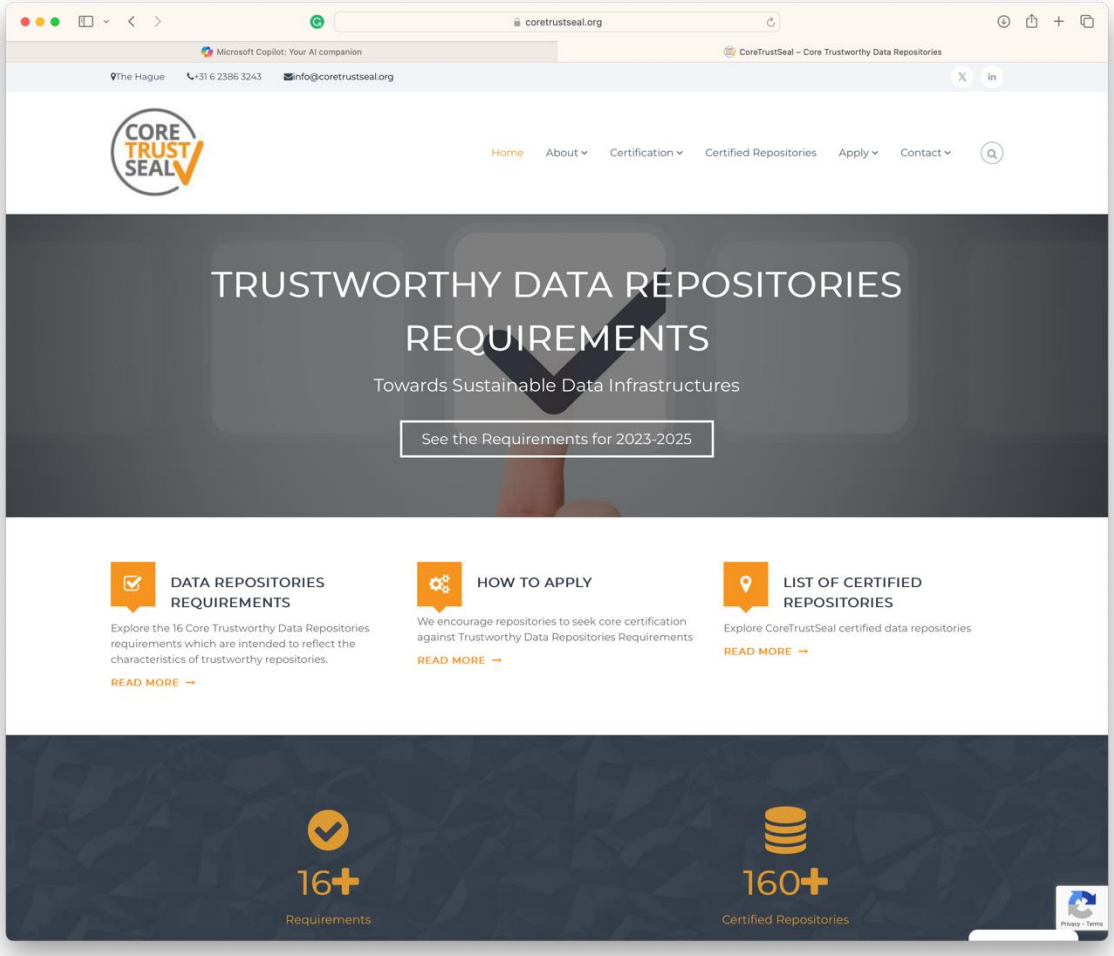

**SIP (Submission Information Package):** Data and metadata received from the producer.
**AIP (Archival Information Package):** Data and metadata organized for long-term preservation within the archive.
**DIP (Dissemination Information Package):** Data and metadata formatted for consumer access.

# CoreTrustSeal

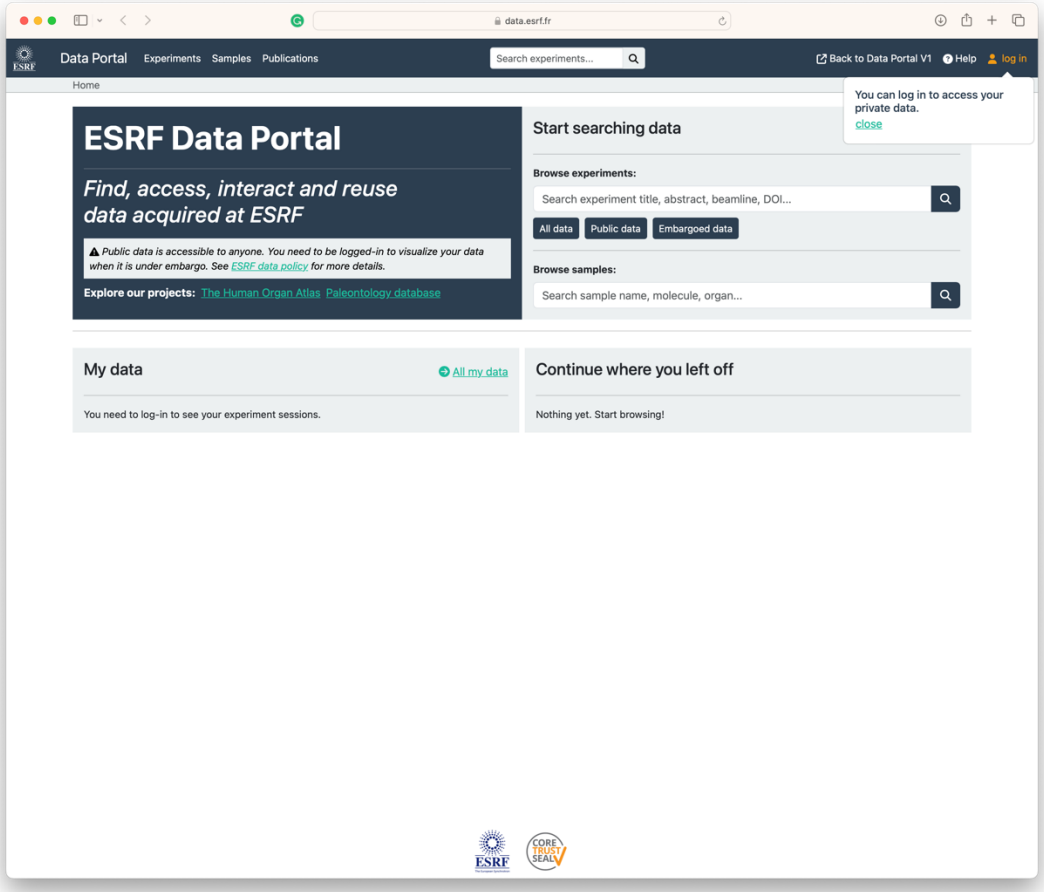
CoreTrustSeal is an international certification for trustworthy data repositories

It defines 16 **requirements**

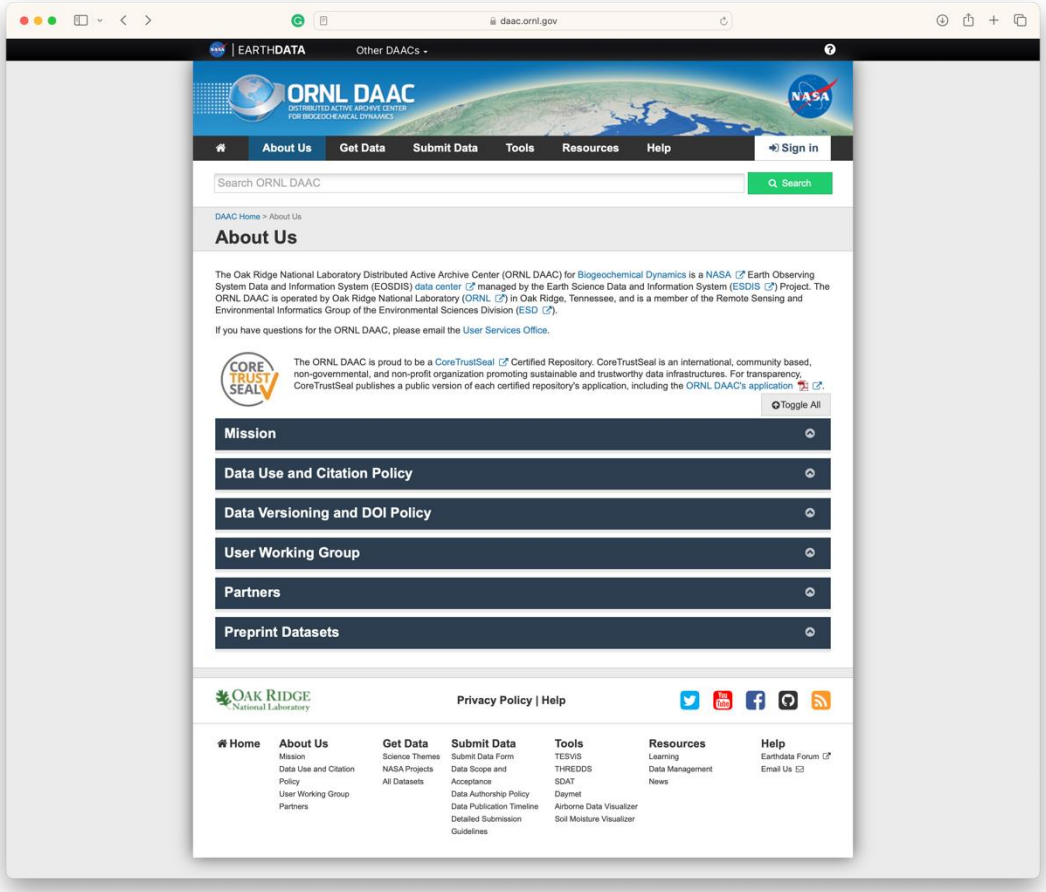
It also includes a self-assessment process and a review by an independent panel of experts



# Example of CoreTrustSeal Repositories

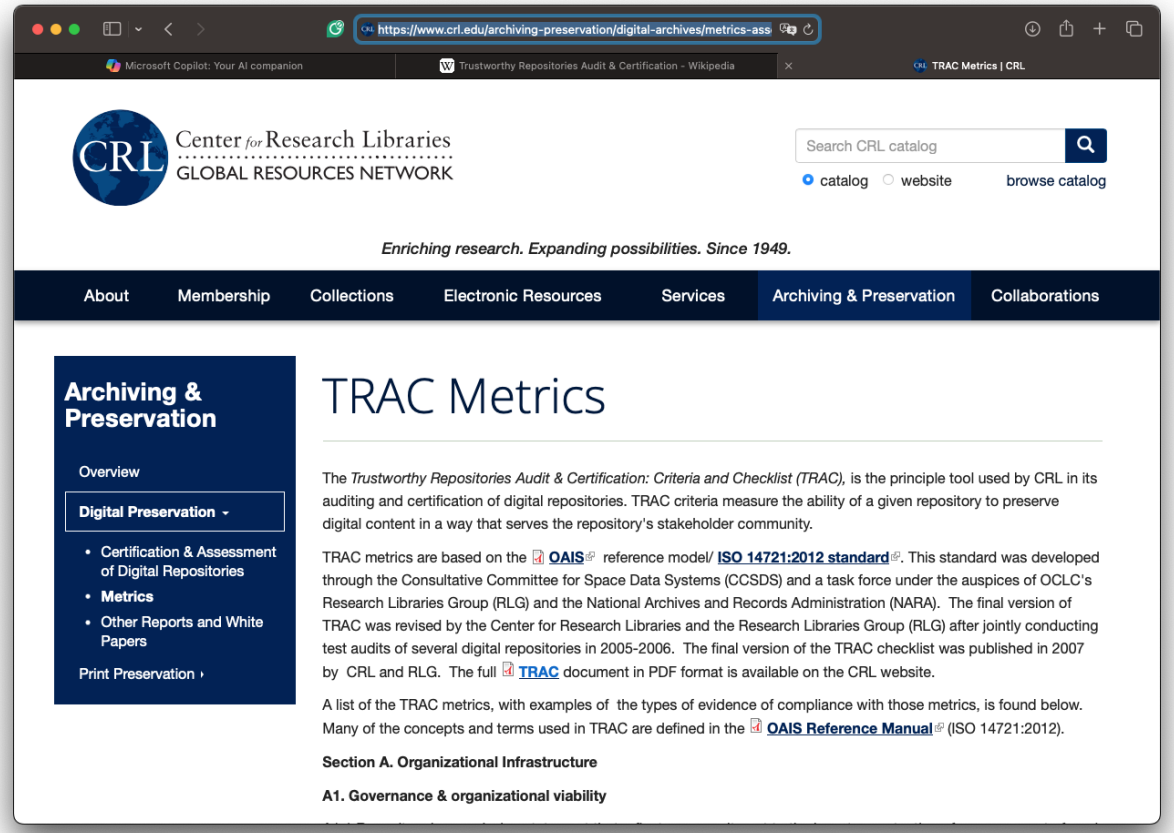



European Synchrotron Radiation Facility (ESRF)




The Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC)

# *Trustworthy Repositories Audit & Certification (TRAC)*



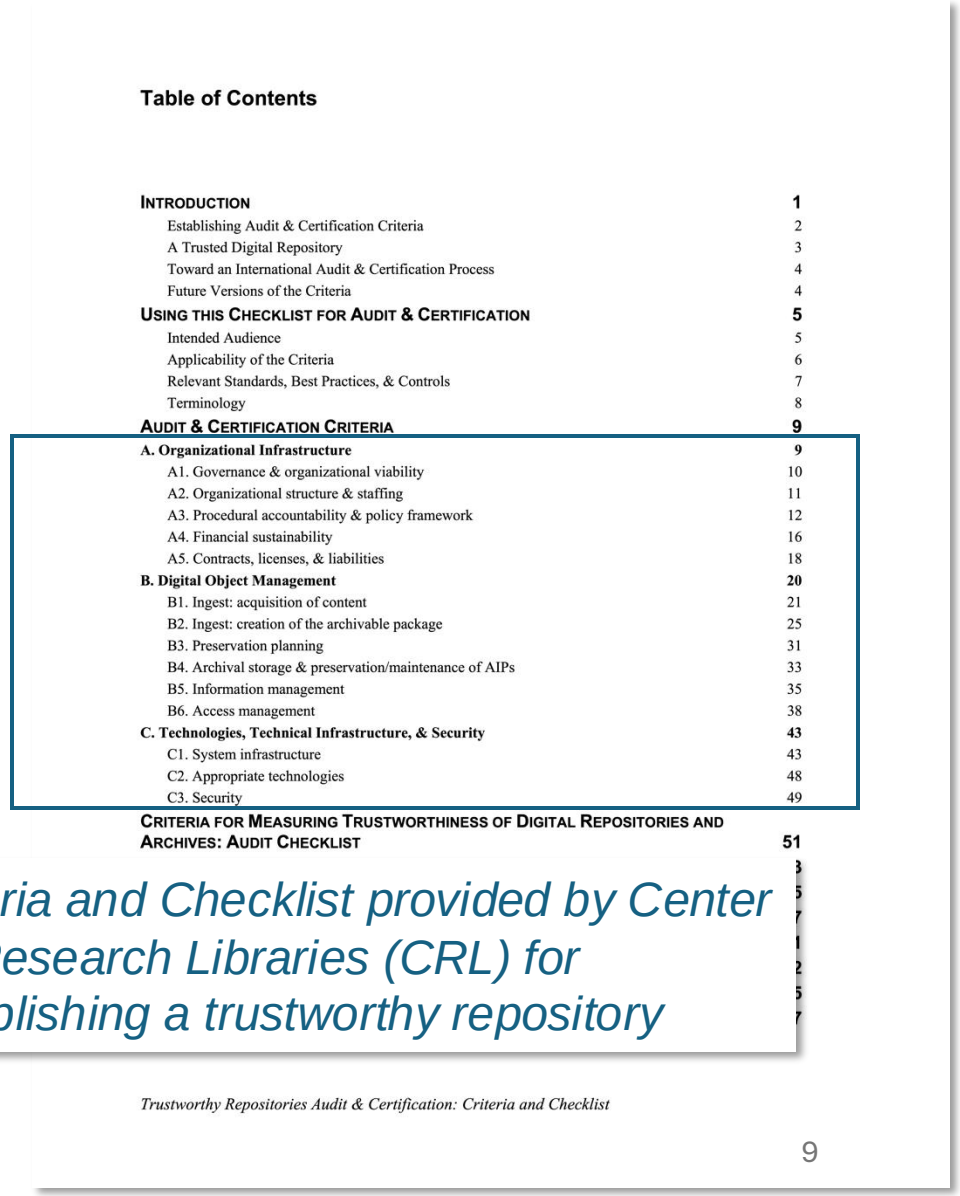



**Table of Contents**



*Trustworthy Repositories Audit & Certification: Criteria and Checklist*

*Criteria and Checklist provided by Center for Research Libraries (CRL) for establishing a trustworthy repository*

A comprehensive and long set of criteria to evaluate the trustworthiness of digital repositories.

# TRAC Metrics Checklist – Example I

**Trustworthy Repositories Audit & Certification: Criteria Checklist**

| | | | |
|---|---|---|---|
| **Organization:** | | **Auditor:** | |
| **Section:** | **A. Organizational Infrastructure** | **Interviewee(s):** | |
| **Aspect:** | **A1. Governance & organizational viability** | | |

| Criterion | Evidence (Documents) Examined | Findings and Obs |
|---|---|---|
| A1.1. Repository has a mission statement that reflects a commitment to the long-term retention of, management of, and access to digital information. | | |
| A1.2. Repository has an appropriate, formal succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope. | | |

**Trustworthy Repositories Audit & Certification: Criteria Checklist**

| | | | | | |
|---|---|---|---|---|---|
| **Organization:** | | **Auditor:** | | **Page** | |
| **Section:** | **B. Digital Object Management** | **Interviewee(s):** | | **Date** | |
| **Aspect:** | **B.1 Ingest: acquisition of content** | | | | |

| Criterion | Evidence (Documents) Examined | Findings and Observations | Result |
|---|---|---|---|
| B1.1. Repository identifies properties it will preserve for digital objects. | | | |
| B1.2. Repository clearly specifies the information that needs to be associated with digital material at the time of its deposit (i.e., SIP). | | | |
| B1.3. Repository has mechanisms to authenticate the source of all materials. | | | |
| B1.4. Repository's ingest process verifies each submitted object (i.e., SIP) for completeness and correctness as specified in B1.2. | | | |
| B1.5. Repository obtains sufficient physical control over the digital objects to preserve them (Ingest: content acquisition). | | | |

# TRAC Metrics Checklist – Example II

**Trustworthy Repositories Audit & Certification: Criteria Checklist**

| | | | | | |
|---|---|---|---|---|---|
| **Organization:** | | **Auditor:** | | **Page** | |
| **Section:** | **C. Technologies, Technical Infrastructure & Security** | **Interviewee(s):** | | **Date** | |
| **Aspect:** | **C1. System Infrastructure** | | | | |

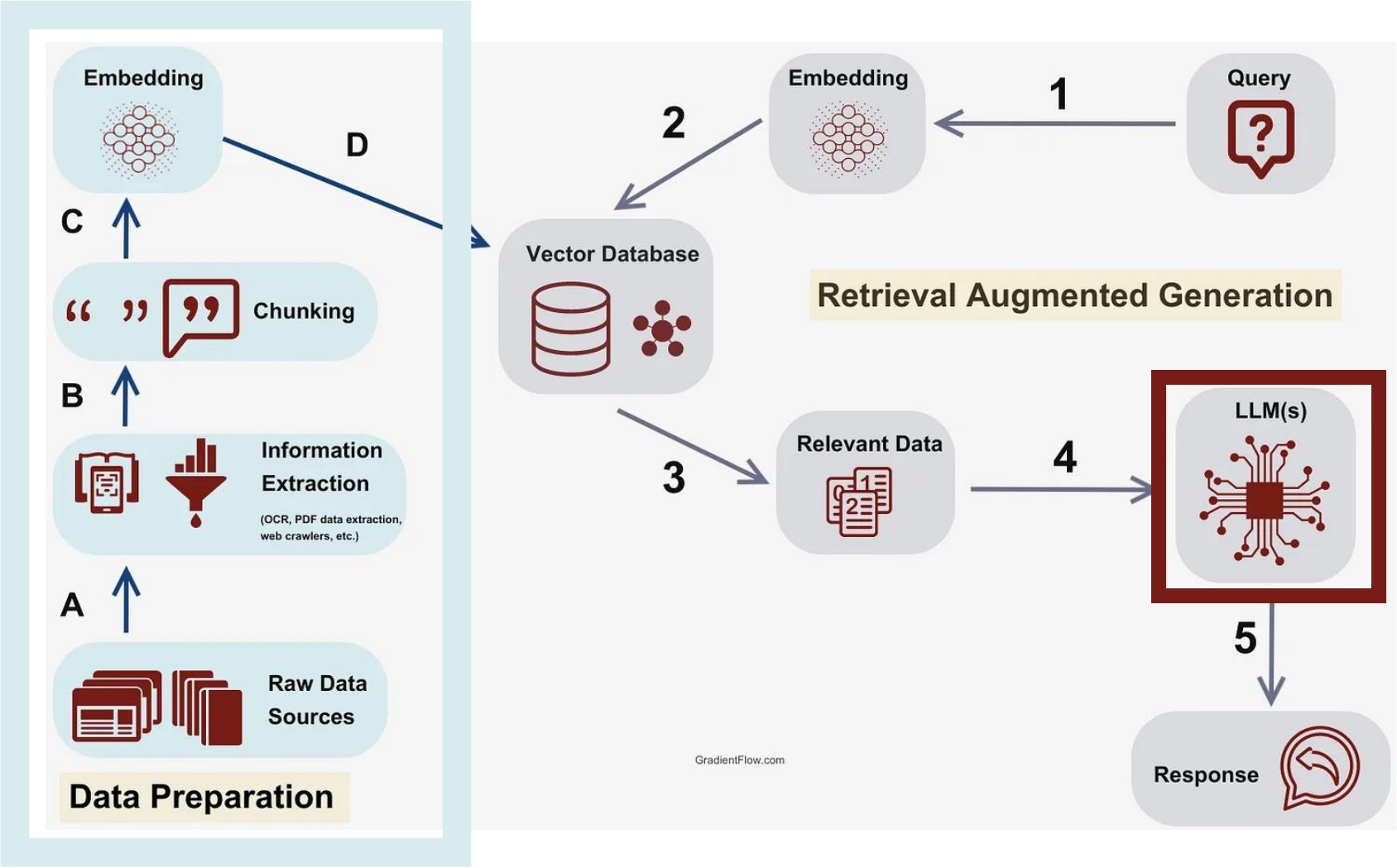
| Criterion | Evidence (Documents) Examined | Findings and Observations | Result |
|---|---|---|---|
| C1.1 Repository functions on well-supported operating systems and other core infrastructural software. | | | |
| C1.2 Repository ensures that it has adequate hardware and software support for backup functionality sufficient for the repository's services and for the data held, e.g., metadata associated with access controls, repository main content. | | | |
| C1.3 Repository manages the number and location of copies of all digital objects. | | | |
| C1.4 Repository has mechanisms in place to ensure any/multiple copies of digital objects are synchronized. | | | |
| C1.5 Repository has effective mechanisms to detect bit corruption or loss. | | | |

# Information Retrieval

- Provide users with a comprehensive understanding of RHIC experiments by combining published data with refined, unpublished data.
- Create a dedicated knowledge-based portal with advanced search capability to centralize and streamline access to RHIC information and data.
  - The technology should be simple to ensure longevity and easy maintenance, while also offering agility to adapt to new solutions as they emerge.
- Ensure the portal supports all RHIC experiments.
  - The portal can be an interface to repositories or a repository, depending on the information type.
  - Federated Identity can be used as a mechanism to adjust access rights.
- Enhance the portal with features powered by Retrieval-Augmented Generation (RAG) and Large Language Models (LLM) for more intuitive and advanced data and information retrieval and exploration.

# Retrieval Augmented Generation (**RAG**)

Domain-specific knowledge

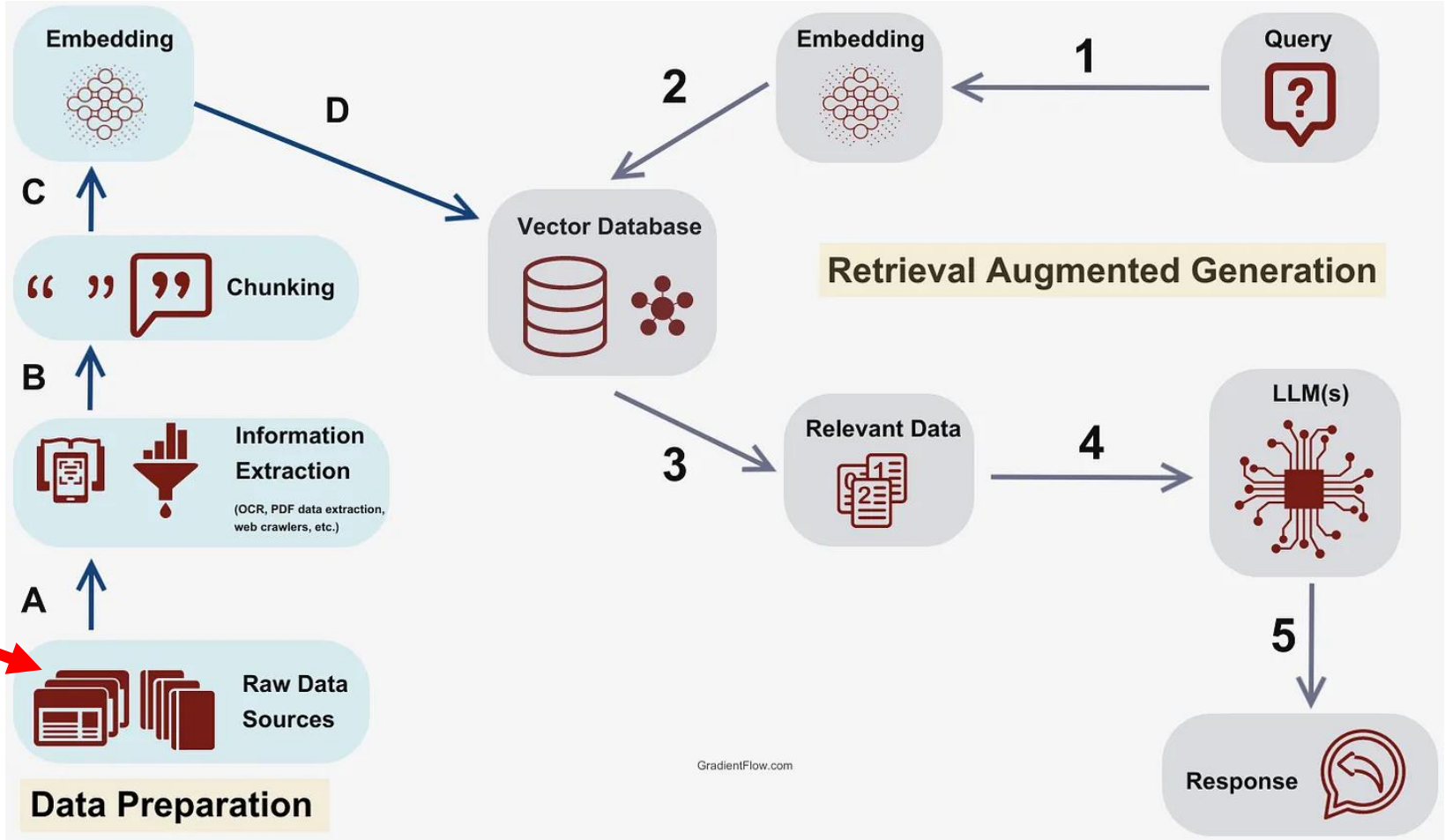



LLM designed for general use



https://gradientflow.substack.com/p/best-practices-in-retrieval-augmented

# Retrieval Augmented Generation (**RAG**)




Published data

Cleaned and validated internal data



https://gradientflow.substack.com/p/best-practices-in-retrieval-augmented

# Thank you

# CoreTrustSeal Requirements by Category

1. **Organizational Infrastructure**: The repository must have a clear mission, adequate staffing, and appropriate policies and procedures.
2. **Digital Object Management**: The repository must manage digital objects effectively, including their ingest, storage, preservation, and access.
3. **Technology and Security**: The repository must use appropriate technology and security measures to protect digital objects and ensure their integrity and accessibility.
4. **Legal and Ethical Compliance**: The repository must comply with relevant legal and ethical standards, including copyright laws and privacy regulations.
5. **Financial Sustainability**: The repository must have a sustainable financial model to ensure its long-term operation.
6. **Transparency and Accountability**: The repository must be transparent in its operations and accountable to its stakeholders.
7. **User Support and Education**: The repository must provide support and education to its users to help them effectively use its services.

These requirements are mandatory and equally weighted, and repositories must provide evidence to demonstrate compliance with each requirement.