



AI Systems Adoption of Unified Research Data Management using oneAPI framework



Digital Research
Alliance of Canada

Alliance de recherche
numérique du Canada

Peter Darveau, P. Eng., CE
Scientific High Performance Computing Engineer
Centre for Advanced Computing
Ottawa, Canada

E-mail: peterd@hexagontechinc.com

About the Author



Peter Darveau P. Eng., has more than 25 years of experience in AI, automation, robotics and has worked with General Electric, General Motors, NASA, FANUC of Japan. He is a professional consulting engineer, oneAPI AI instructor and practitioner. He has published papers relating to automation, High-Performance Computing (HPC) and AI models and has been cited internationally. Speaker at events for GE and at national level for OSPE, SME and Digital Research Alliance of Canada. Active with HPC in research and AI startups and as well as within the engineering and francophone communities in Toronto.



Peter Darveau P. Eng., CED, 6sigma
GB, MBA

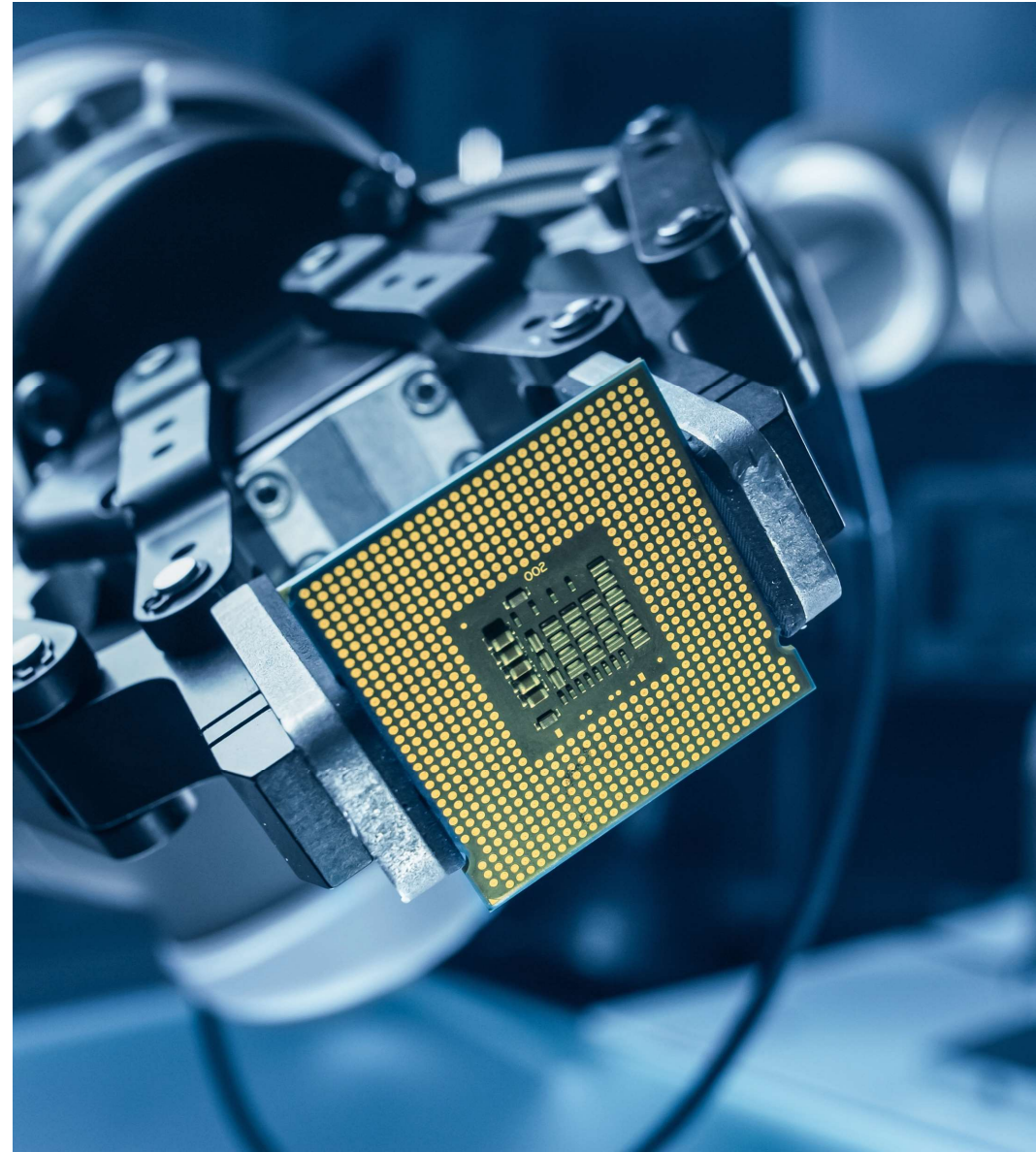
| RESEARCH IT & ENGINEERING
PROFESSIONAL | MACHINE LEARNING | INT...



Section 01

RDM and AI Modeling - Problem Background

- Data Curation mostly limited to:
 - Data Repo Descriptive Attributes
 - File naming conventions
- Tracking, tracing and verifying to past research – reproducibility
- Funding qualification requirements
- More AI applications are moving to the Cloud



Evaluated through use cases, plenary sessions and discussions, dataset attributes are:

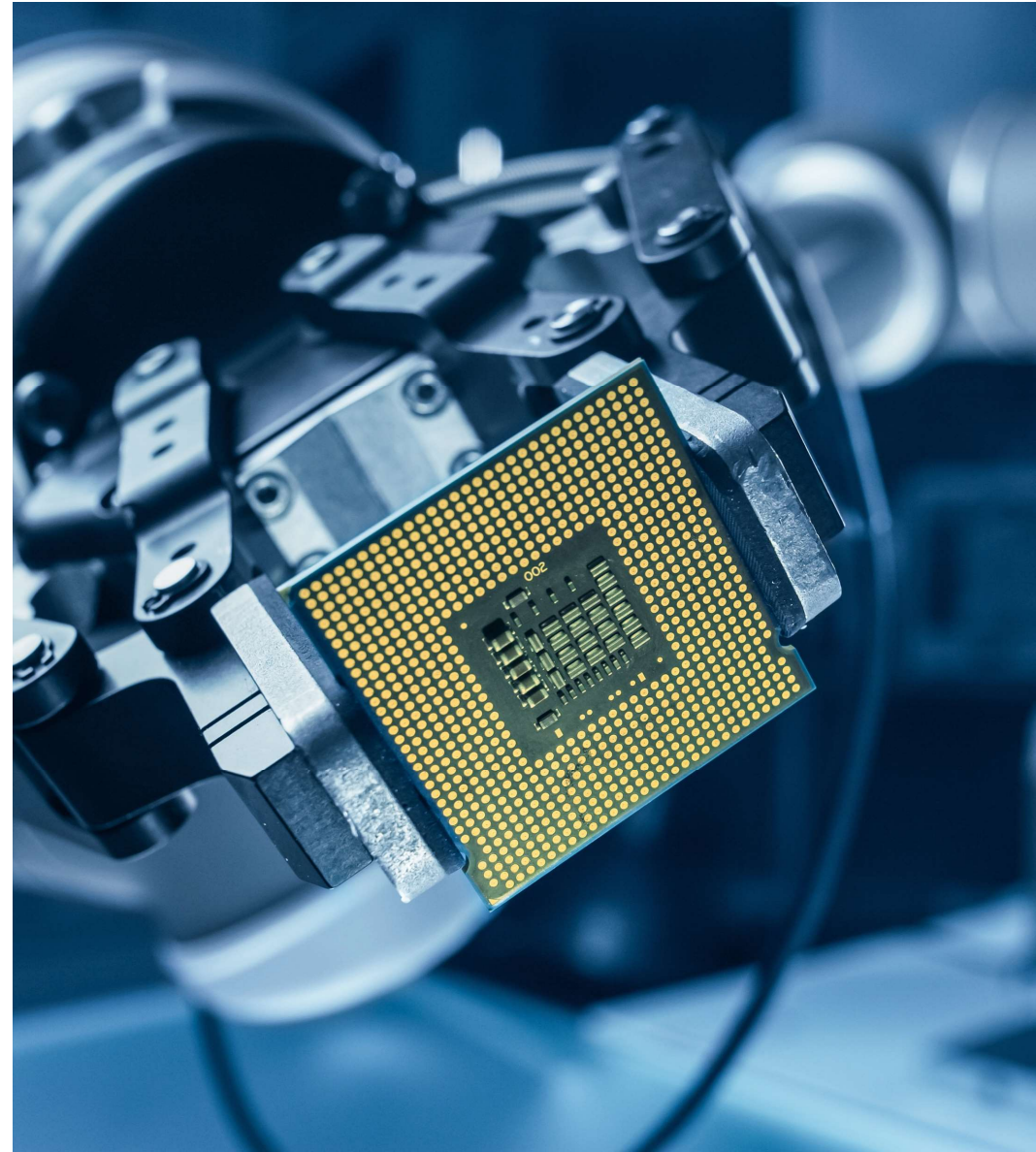
- Repository Name
- URL
- Country
- Language
- Organization
- Contact
- Description
- Research Area
- Persistent Identifiers
- Machine Interoperability
- **Metadata**
- Terms of Deposit
- Terms of Access
- Dataset Use License
- Certification
- Preservation



Section 02

Trends in Curation - Metadata

- Datasets increasingly integrated into the research project. An example [New Integrations and Tools - Fall 2023 - XWiki \(scholarsportal.info\)](#)
- Achieved via the use of metadata format in a configuration file
- Metadata standards are starting to be recognized as an important ingredient to improve data discoverability – especially for restricted or sensitive data
- Approaching a more algorithmic organization



Section 03

Humanities & Medicine - Challenges

- AI computing trends are moving to the Cloud
- Projects are incorporating IoT data which needs multi-user sites scaling across the globe
- Secure computing focus on the Network, Storage – Not the silicon
- The amount of restricted and sensitive data is growing fast and 75% of it is video data requiring high performance computing



What is restricted and sensitive data?

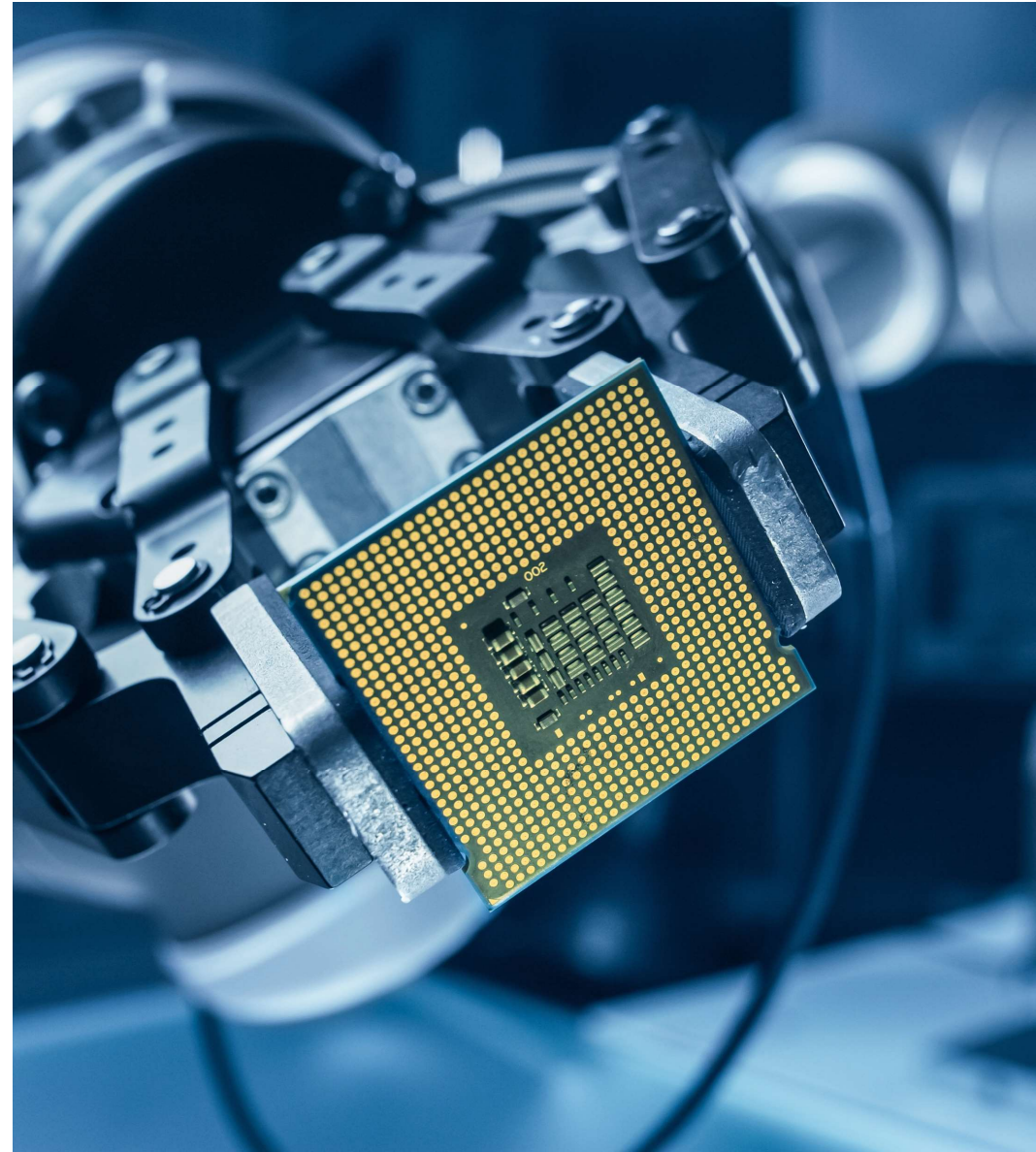
- Personally Identifiable Information (PII)
- Personal Health Information (PHI)
- Commercial data
- Demographic data
- Indigenous data
- Genealogical data
- Educational records
- Criminal records
- Financial Records
- Geographic information
 - (e.g., endangered species data)
- Proprietary data
- Research data about people
- Climate change data
- Cultural heritage data
- Community data
- Voice recordings
- Data with national security considerations
- Social media data
- Embargoed data
- Data “deemed confidential”
- Government administration data
- Any data that could cause harm



Section 01

RDM and AI Modeling - Problem Keeps Moving

- Cloud Regulation and Directive concerning cybersecurity countermeasures
 - Seen more as a Network problem and NOT a computing problem
 - Attack surface is complex and constantly changing, along with the technology used to mitigate attacks, regulation tends to be technology neutral
 - Lack of standards at the computing device level



To Summarize ...

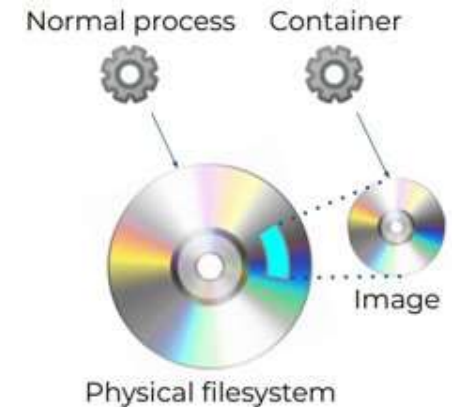
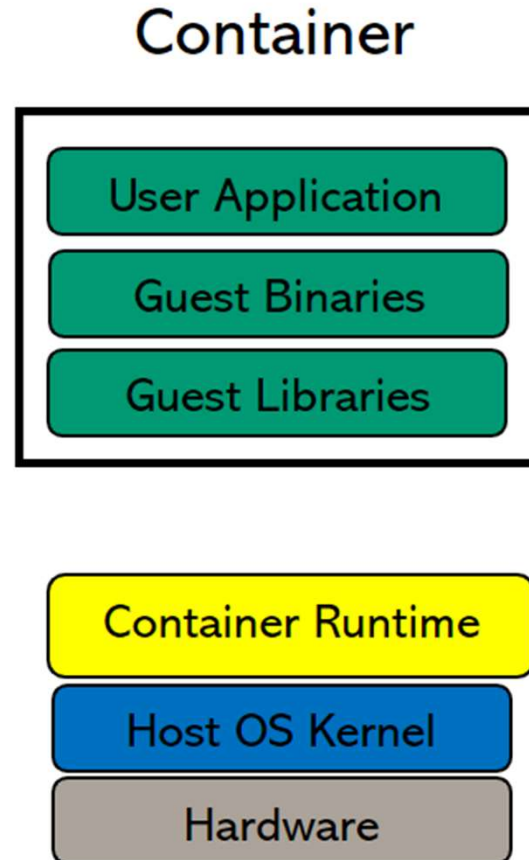
Proper data handling and integration into the project is key to addressing the data curation problem.

But that solution will be very short lived without software discipline that unifies and secures the computational resources.

Section 05

Computing Environment - Containerization

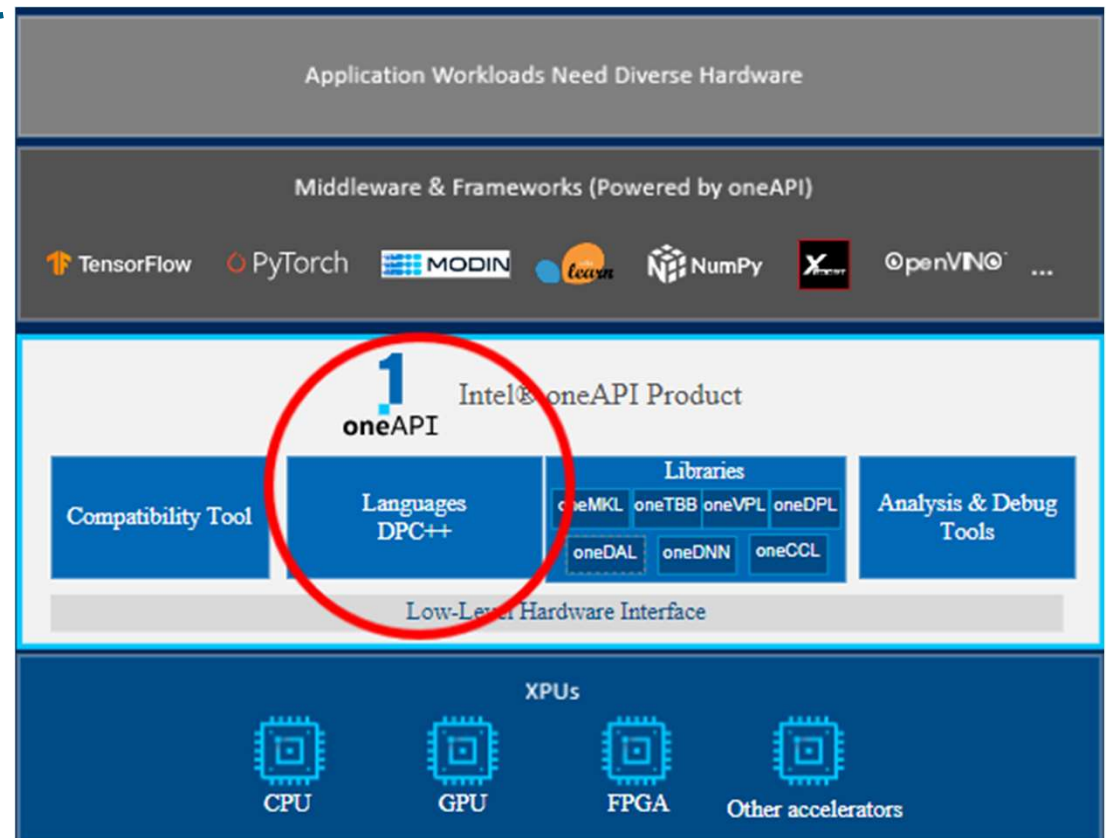
- A container is a process that has its own view on local resources
- In the picture, the container sees the image of the physical filesystem
- Reproducibility
 - Largely unaffected by changes to the cluster environments
- Security
 - Isolated environments to secure development or runtime



Section 06

Research Development Software

- Unified Accelerator Standard – oneAPI
 - CPU, GPU, XPU (ASIC)
- DPC++
 - unifies programming of central processing unit (CPU, scalar computation), graphics processing unit (GPU, vector computation), artificial-intelligence accelerator (AI, matrix or tensor) and field programmable gate array (FPGA, spatial computation)
 - Device code can be parallelized
 - Public and private access modifiers as in C++ Object-Oriented Programming
 - Configuration file .yml can be used to declare validation process. This file contains the metadata containing computational parameters for unified and secure computing



Example Metadata file

```
models:  
  - name: image_model  
    launchers:  
      - framework: dlsdk  
        tags:  
          - FP32  
        model: ./data/public/model/FP32/image_model.xml  
        weights: ./data/public/model/FP32/image_model.bin  
        adapter: ssd  
        device: CPU # or GPU, GPU.1, GPU.2  
  
  datasets:  
    - name: image_model_detection_91_classes  
      data_source: ./data/datasets/model_val_data/val2024  
      annotation_conversion:  
        annotation_file: ./data/datasets/model/annotations/instances_val2024.json  
        access_mod: public # or private  
        has_background: True  
        use_full_label_map: True  
        converter: img_detection  
      preprocessing:  
        - type: resize  
          size: 300  
      postprocessing:  
        - type: resize_prediction_boxes  
      metrics:  
        - type: precision # or accuracy, top_k, F1_score, etc.
```




Section 07


Research Governance Model


- Where does this framework fit in the governance?

Role	Governance				
Researcher	Data		Infrastructure		
	Laws/Regulations	Standard Ontology	Policies	Lifecycle Management	Sustainability
Software Developer	Data Competencies				
	Data Management			Data Guardrails	
	Working with Data	FAIR implementations	Optimization / Performance	Data methods	Data Structure
Data Manager / Custodian	Research Software Structured for Data Flow, Security, Verification and Validation				
	Encapsulation	Abstraction	Standards	Data Preservation / Integrity	Containment

This paper







Section 08

Future Work

- Data is currency in an AI economy.
- Eliminating barriers to data access is a powerful catalyst for innovation.
- Access to computing power and AI technologies are also a necessary ingredient ...

This is best made possible with ...

A unified standard around computing with sensitive data in a trusted and secure way will become a growing part of the data governance model



Some of my recent work on this topic

CAC Machine Learning series incorporates training on RDM best practices

- EN [Machine Learning \(uottawa-it-research-teaching.github.io\)](https://uottawa-it-research-teaching.github.io)
- FR [Apprentissage-machine \(uottawa-it-research-teaching.github.io\)](https://uottawa-it-research-teaching.github.io)
- Research IT
[Recherche uO Research: Peter Darveau](#)



Peter Darveau P. Eng., CED, 6sigma
GB, MBA

| RESEARCH IT & ENGINEERING
PROFESSIONAL | MACHINE LEARNING | INT...

