Tracking and Sharing of Data Provenance for Scientific Workflows

PI: V. Judson Harward (MIT)

Liang Yu (NUS), Philip H. Bailey (MIT), Chen-chieh Feng (NUS), James Hardison (MIT), Hanna Kurniawati (SMART)

1. Motivation

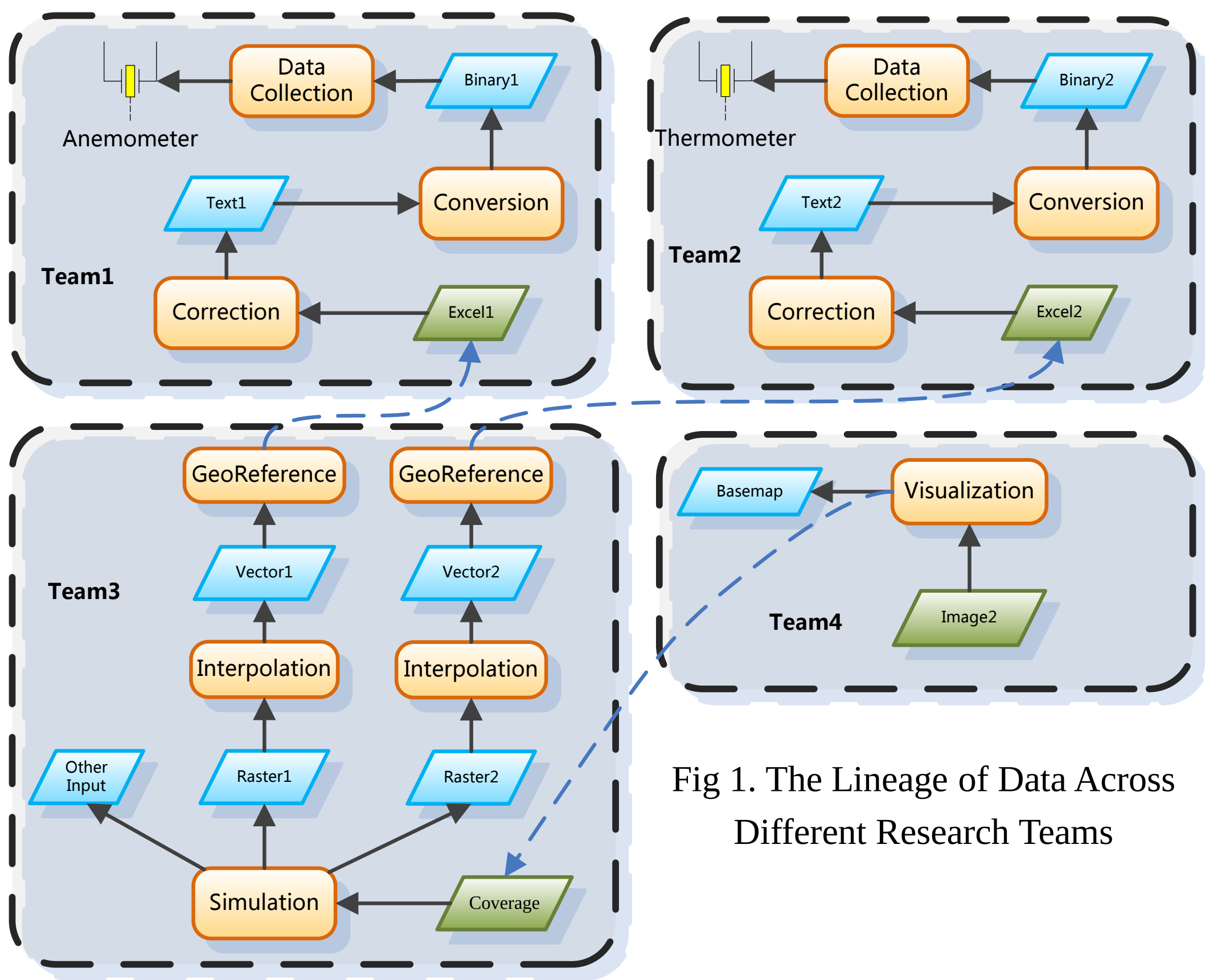


Fig 1. The Lineage of Data Across Different Research Teams

Cyberinfrastructure provides advanced services for multiple research organizations to share data with each other. A dataset might be used for various purposes other than its initial intended use. It is crucial for users to know where a dataset comes from and how it was generated before they make use of it. The concept of provenance is introduced as linkage between objects.

With provenance information, data users can trace the history of a dataset by analyzing each of its causal elements, evaluate the data quality by using error propagation models, or reproduce the dataset while sometimes changing the inputs or parameters. In a word, data provenance can help users adapt the data properly to their own application.

2. System Framework

Two key issues related to a provenance system are capture and management. In our current prototype, we use Linux-based shell scripts to drive a scientific workflow which generates new datasets and automatically records lineage information. We also considered the tools for manually editing provenance. For exchanging provenance information, the Open Provenance Model (OPM) is adopted as both the exchange format (OPM XML Spec) and the basis of the conceptual model for management (OPM Ontology). The provenance data are stored in a Resource Description Framework (RDF)-enabled database and can be accessed through web services. The triple structure of RDF is very flexible and well suited for building scalable queries.

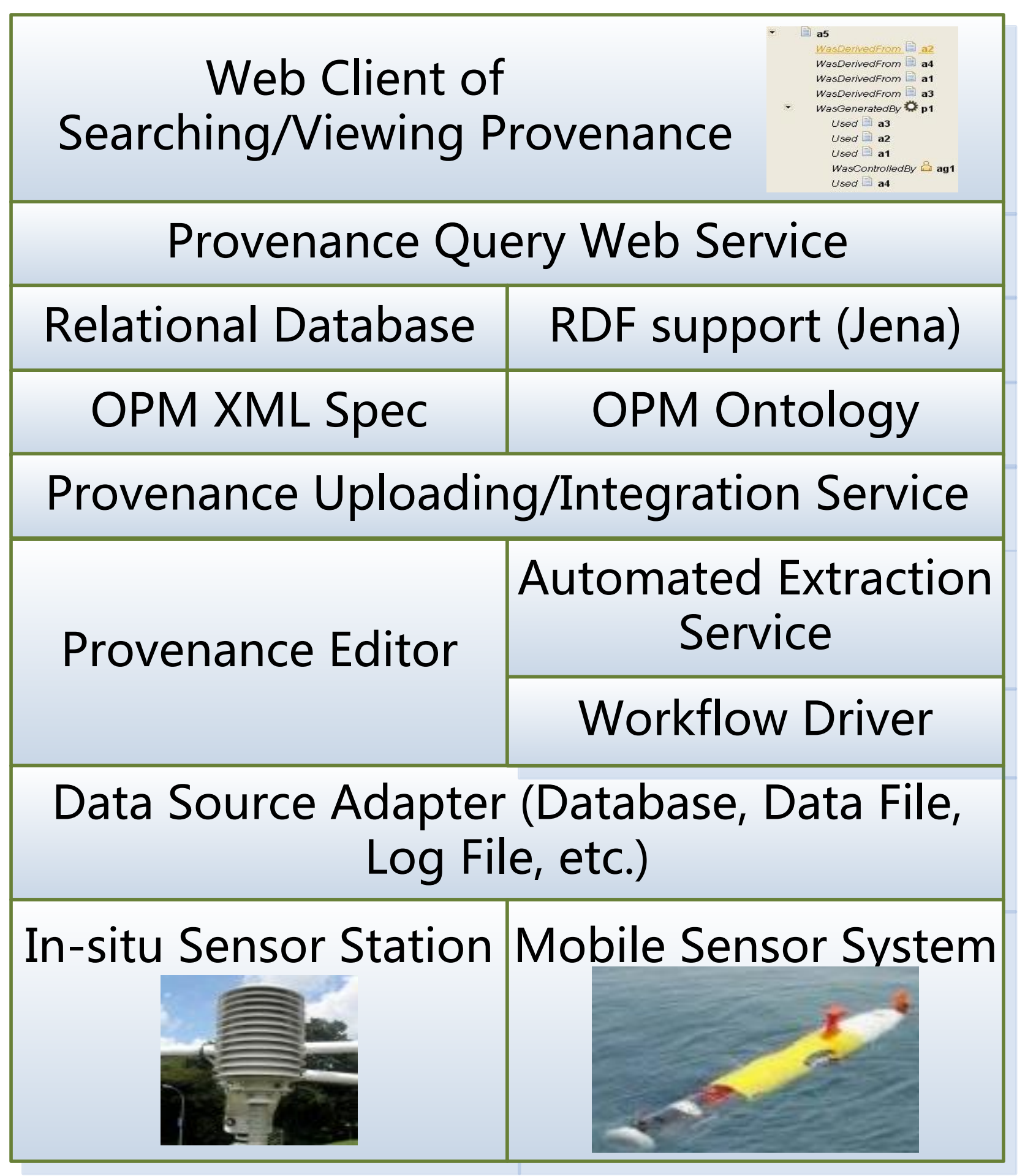
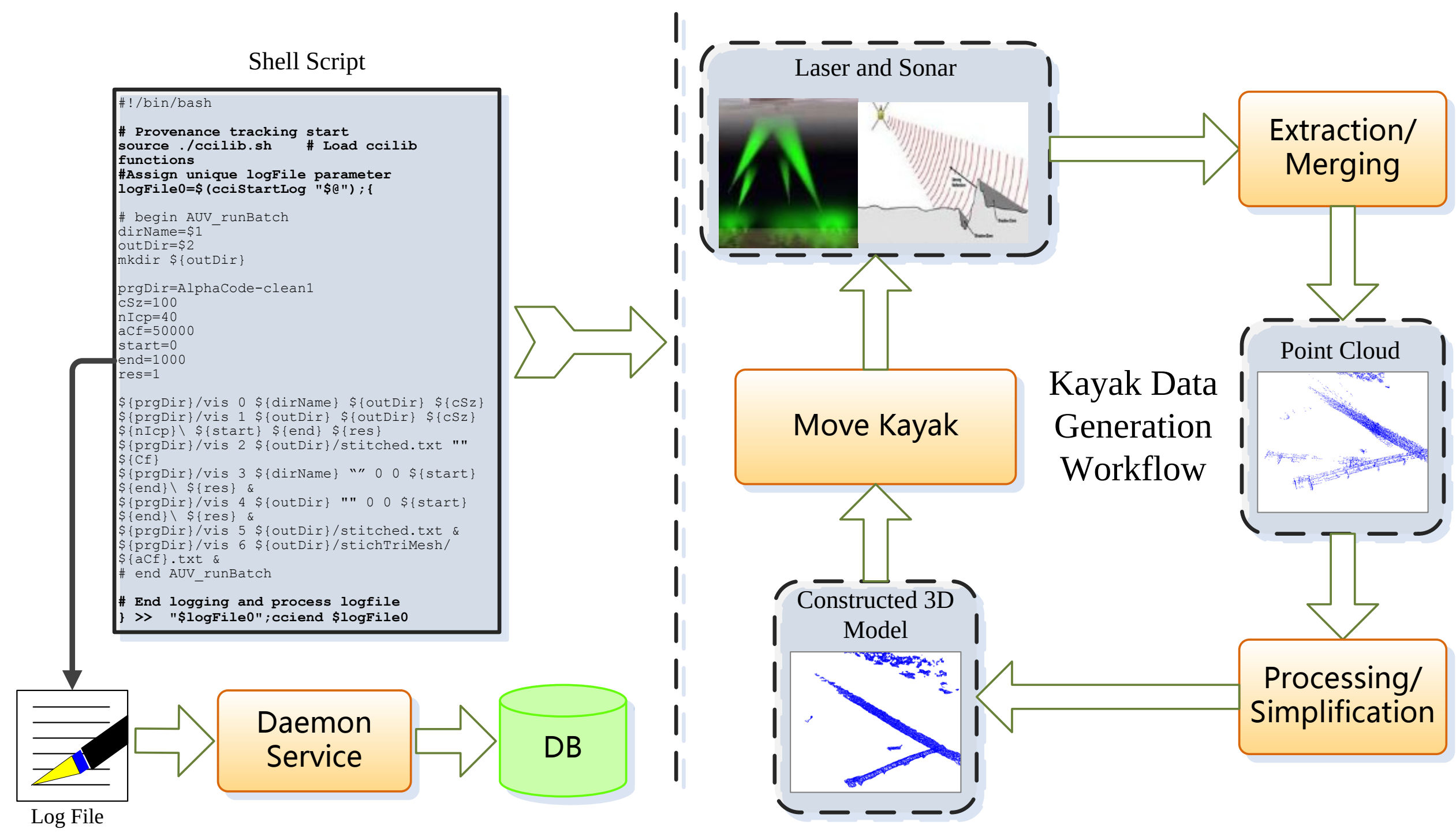

Web Client of Searching/Viewing Provenance
Provenance Query Web Service
Relational Database
RDF support (Jena)
OPM XML Spec
OPM Ontology
Provenance Uploading/Integration Service
Provenance Editor
Automated Extraction Service
Workflow Driver
Data Source Adapter (Database, Data File, Log File, etc.)
In-situ Sensor Station
Mobile Sensor System

Fig 2. Frame work of Provenance System

3. Shell script Assisted Provenance Capturing



```
#!/bin/bash

# Provenance tracking start
source ./ccilib.sh     # Load ccilib
functions
#Assign unique logFile parameter
logFile0=$(cciStartLog "$@");{

# begin AUV_runBatch
dirName=$1
outDir=$2
mkdir ${outDir}

prgDir=AlphaCode-clean1
cSz=100
nIcp=40
aCf=50000
start=0
end=1000
res=1

${prgDir}/vis 0 ${dirName} ${outDir} ${cSz}
${prgDir}/vis 1 ${outDir} ${outDir} ${cSz}
${nIcp}\ ${start} ${end} ${res}
${prgDir}/vis 2 ${outDir}/stitched.txt ""
${Cf}
${prgDir}/vis 3 ${dirName} “” 0 0 ${start}
${end}\ ${res} &
${prgDir}/vis 4 ${outDir} "" 0 0 ${start}
${end}\ ${res} &
${prgDir}/vis 5 ${outDir}/stitched.txt &
${prgDir}/vis 6 ${outDir}/stichTriMesh/
${aCf}.txt &
# end AUV_runBatch

# End logging and process logfile
} >>  "$logFile0";cciend $logFile0
```

Fig 3. A shell script drives the continuous workflow and records the provenance information in a log file simultaneously

The figure on the right side depicts a recursive scientific workflow which generates a large number of datasets. These datasets are linked to each other in a hierarchical structure. When there is an error identified, researchers have to trace and analyze all the inputs and parameters. This workflow is driven by shell scripts, one is shown on the left side. CCI functions have been added to the script which records the provenance information into a log file and triggers a daemon service used to extract information from the log file and import it into the database. This is only one example of the various scientific workflows in CENSAM. Others can be driven by different workflow engines. The key point is, whenever a single process finishes, its provenance information is recorded, which includes inputs, outputs, starting and ending time, etc.

4. Provenance Management with OPM

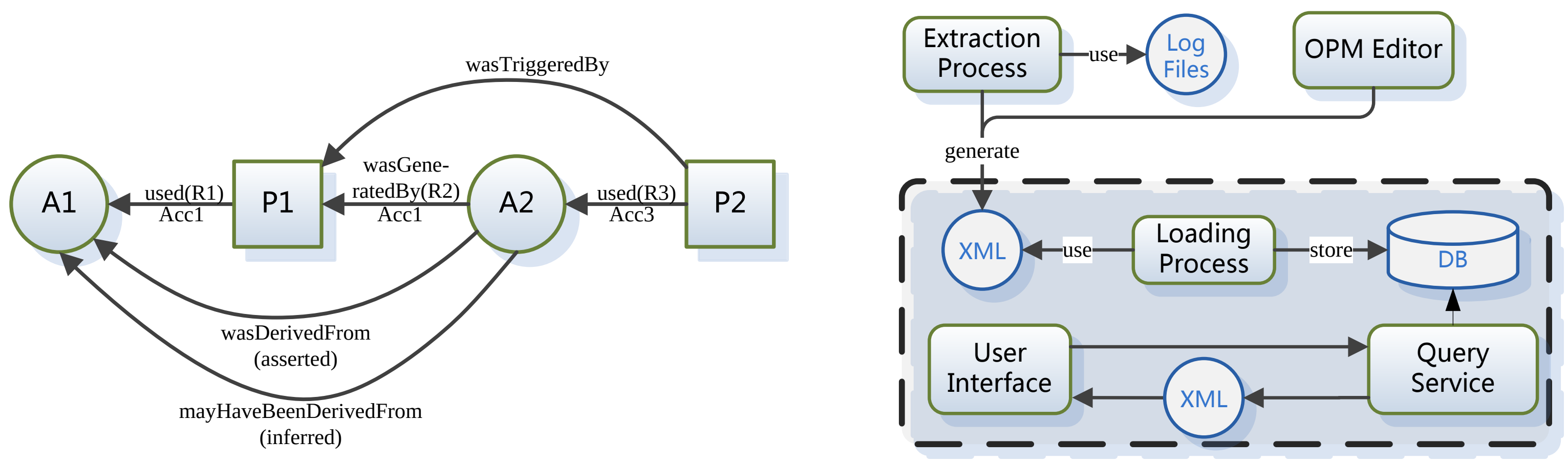


Fig 4. (a) OPM Elements. (b) The “Import” and “Export” Functions of Provenance System

OPM provides a framework of describing provenance. It defines three basic elements, *Artifact*, *Process* and *Agent*, along with a set of relations between them, all of which are concepts in the OPM ontology. The OPM ontology can also be used as a conceptual model for storing the provenance with the support of an RDF database. The OPM XML Specification provides a standard for provenance exchange, which is used for import and export of the provenance system.

5. Provenance Integration

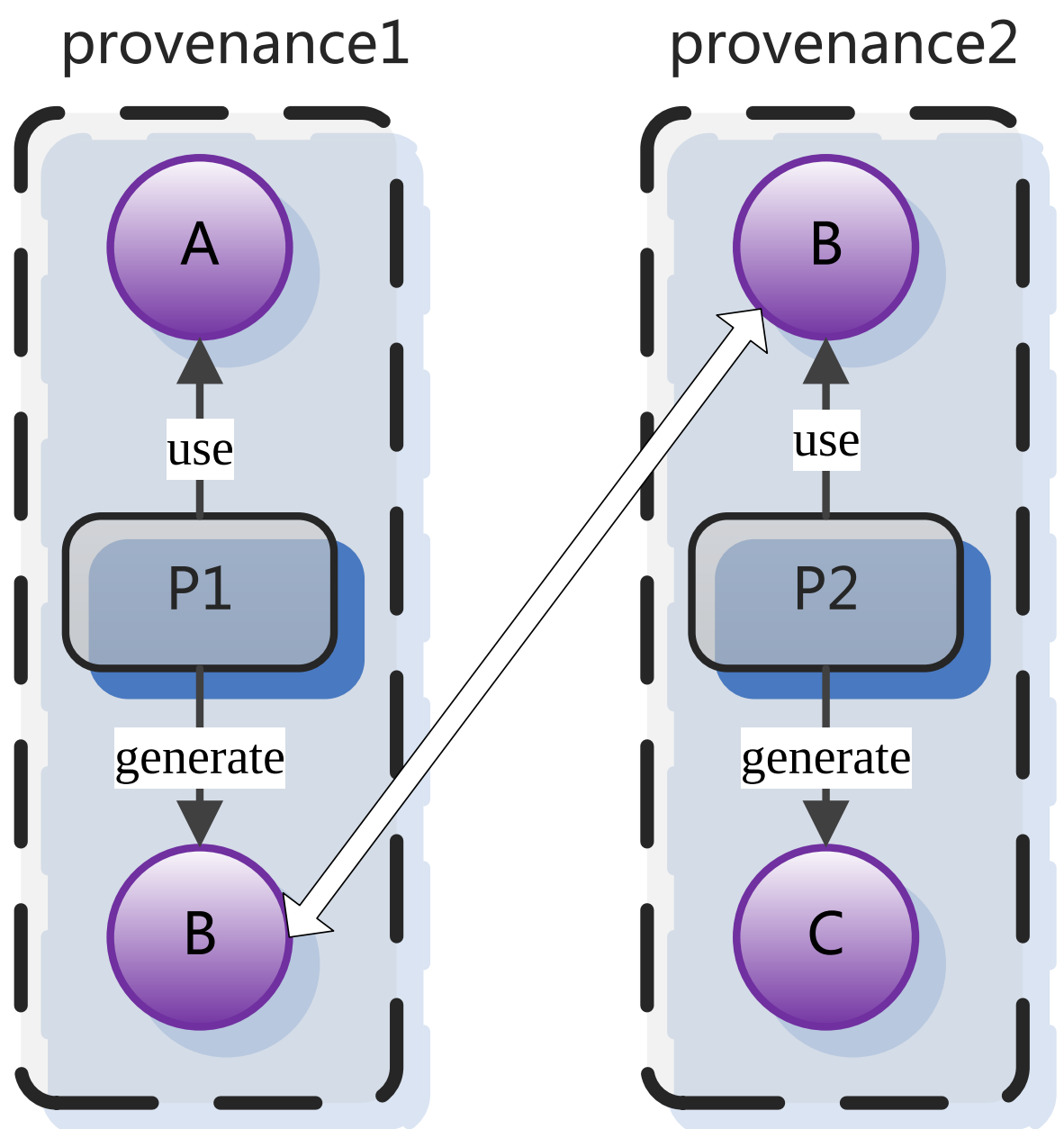


Fig 5. Connection between two pieces of provenance

Provenance of a workflow is usually recorded as a package (or a *Graph* in OPM). However, elements from different packages might have various types of connections, e.g., a process uses a dataset generated from another process and each generates a provenance file. In order to integrate these provenance data, a Global Unique Identifier (GUID) mechanism is proposed to uniquely recognize the objects. Two mechanisms are taken into consideration: URL serves as GUID for published resources while a digest algorithm can serve for file based datasets.

6. SOA based System Integration

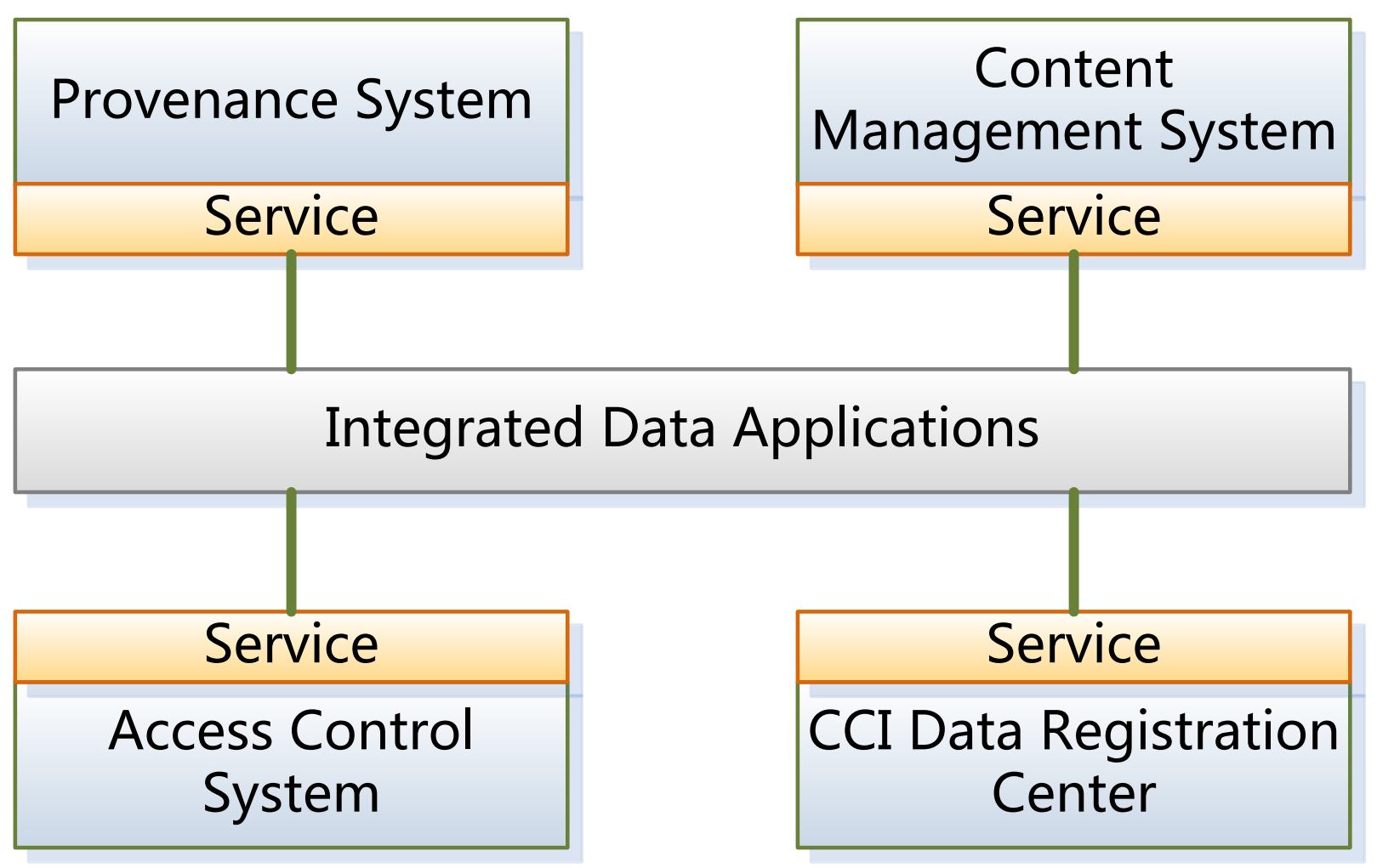

Provenance System
Service
Content Management System
Service
Integrated Data Applications
Service
Access Control System
Service
CCI Data Registration Center

Fig 6. System Integration in CyberInfrastructure via Web Services

Four main systems currently in the CyberInfrastructure are Content Management System, Access Control System, Data Registration System and Provenance System. Every system functions individually but are loosely dependent on each other. By virtue of the Service Oriented Architecture (SOA), the provenance system can locate the real data in the data registration system or content management system by using a global identifier, and access the data source if the operation is approved by the access control system. Moreover, any clients can query the provenance by providing an GUID.

7. Future Work

1. As additional requirements arise to handle complicated processing logic, there is a need to investigate more generic workflow tools which can drive scientific workflow in different environments and are customizable for capturing provenance.

2. The OPM should be extended to include more detailed provenance such as parameters and scientific models. To make it consistent, the OPM ontology should be extended to include or import more concepts from domain ontologies.

3. New tools can be developed by using provenance. For example, when an error is discovered in a data file, a tool can search the content management system to find out all the results that could have been affected by it; or by analyzing two similar workflows, a tool tell what inputs cause the difference between their outputs.

Acknowledgements

The research described in this project was funded in whole or in part by the Singapore National Research Foundation (NRF) through the Singapore-MIT Alliance for Research and Technology (SMART) Center for Environmental Sensing and Modeling (CENSAM).