Provenance for seismology

Alessandro Spinuso

University of Edinburgh (Uk)

KNMI (Royal Institute of Meteorology - NL)

Outline

Seismological Use cases

Workflow system and provenance aware components

Roles of Provenance

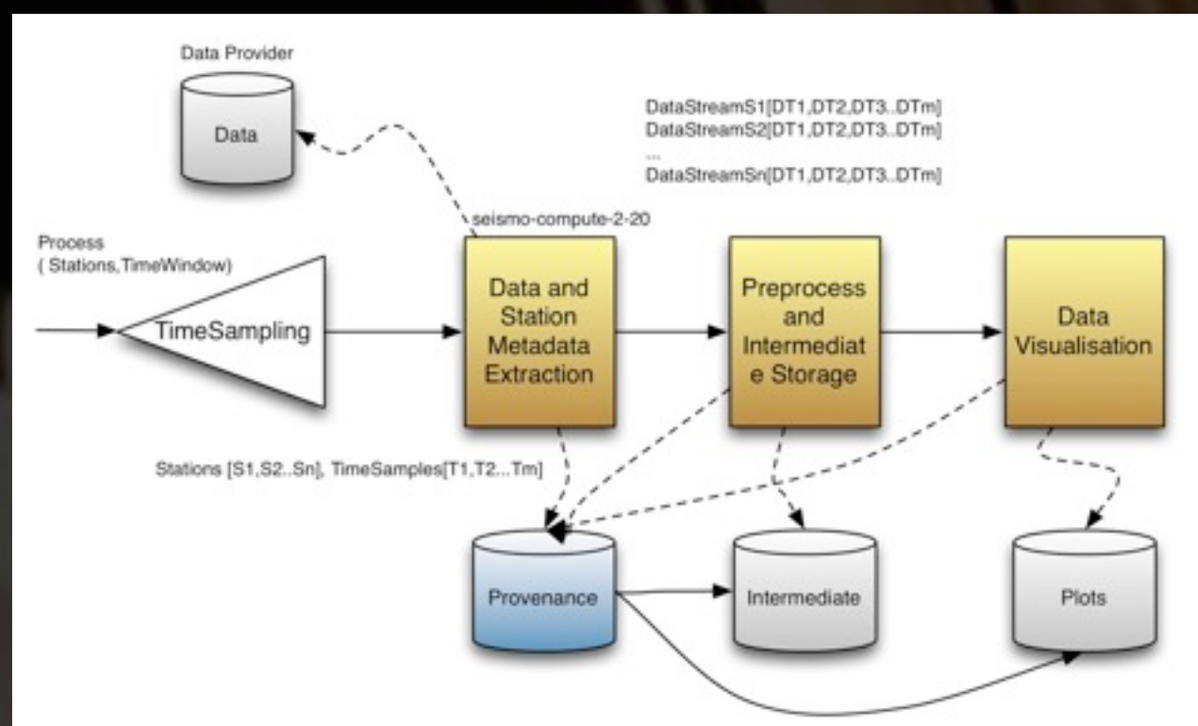
Harvesting Challenges and Strategy in DISPEL

Seismological Pipelines

Processing Chains are common practice in

Seismic Interferometry [G. D. Bensen et al. 2007]

Forward Modelling [A. Fichter et al. 2008]

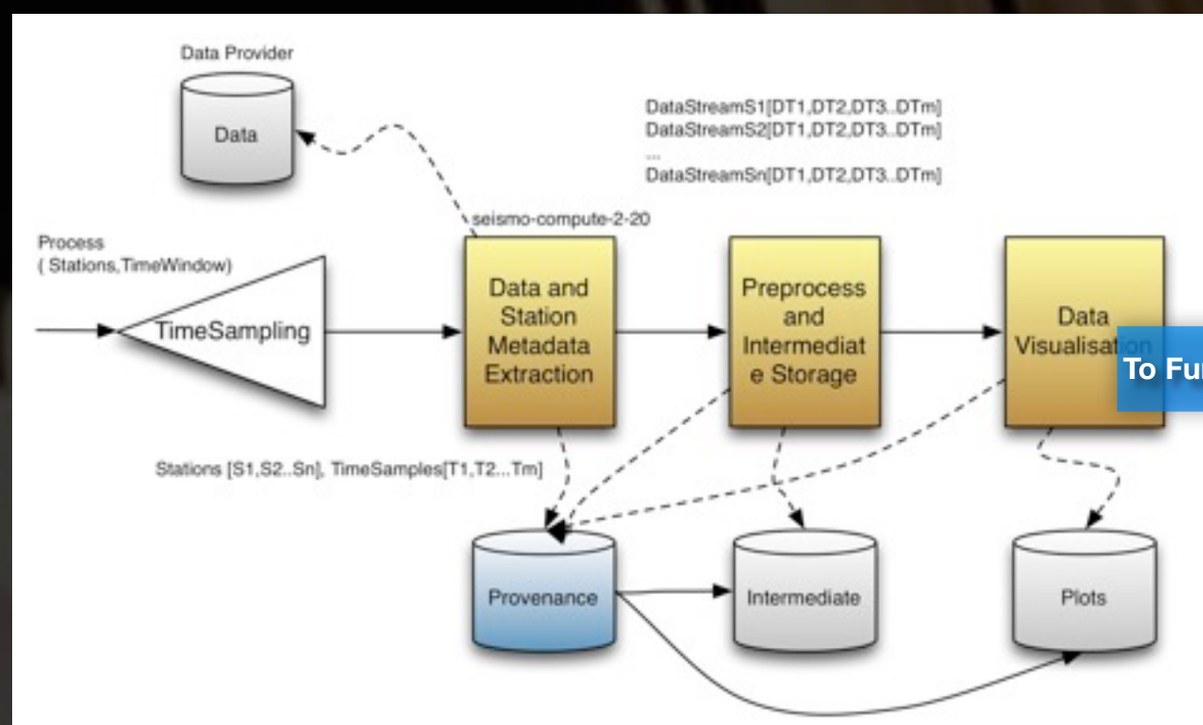


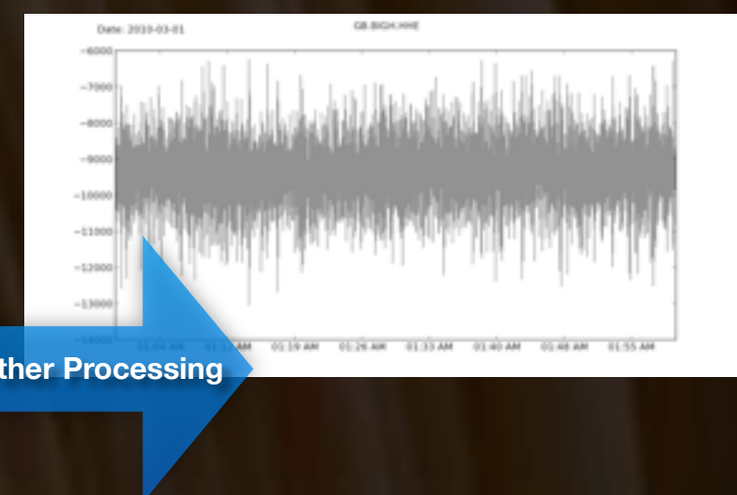
Seismological Pipelines

Processing Chains are common practice in

Seismic Interferometry [G. D. Bensen et al. 2007]

Forward Modelling [A. Fichter et al. 2008]



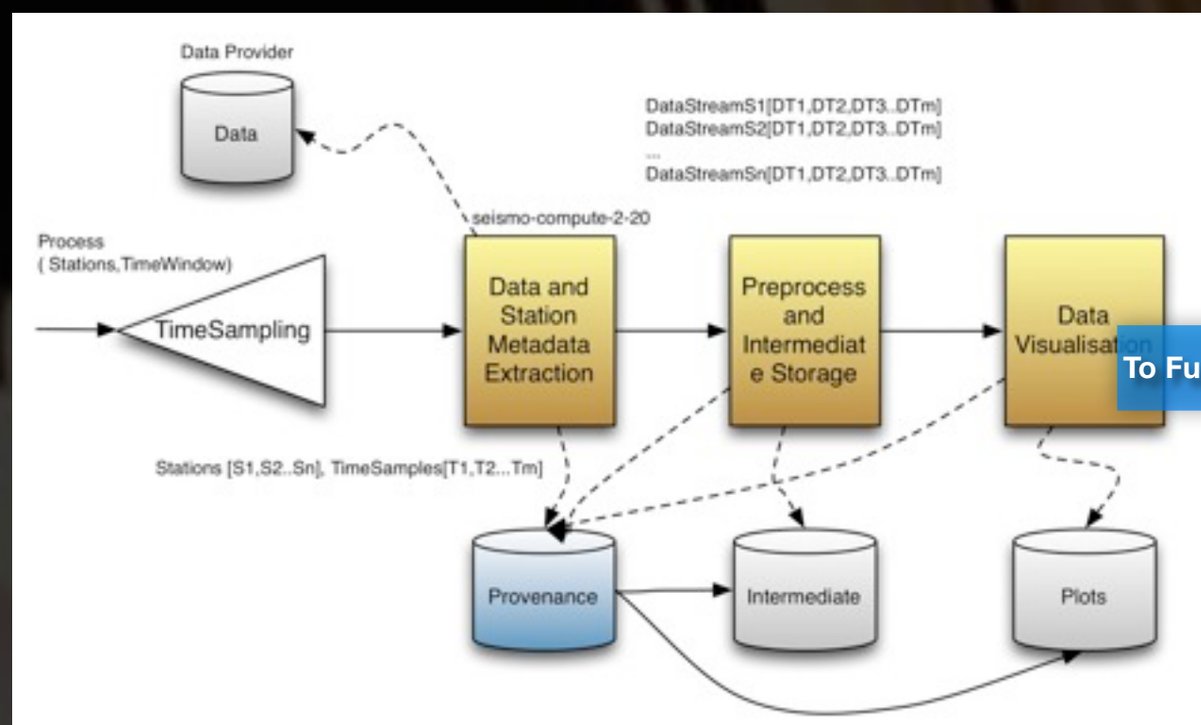


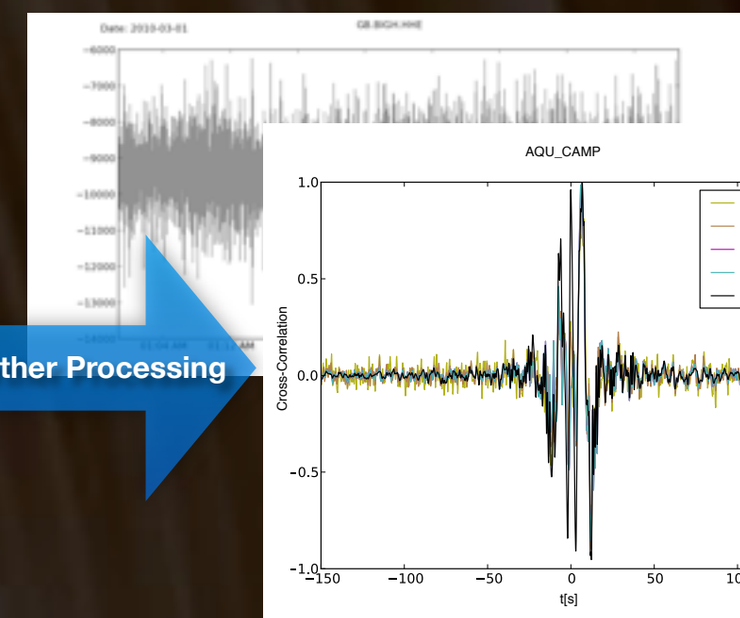
Seismological Pipelines

Processing Chains are common practice in

Seismic Interferometry [G. D. Bensen et al. 2007]

Forward Modelling [A. Fichter et al. 2008]





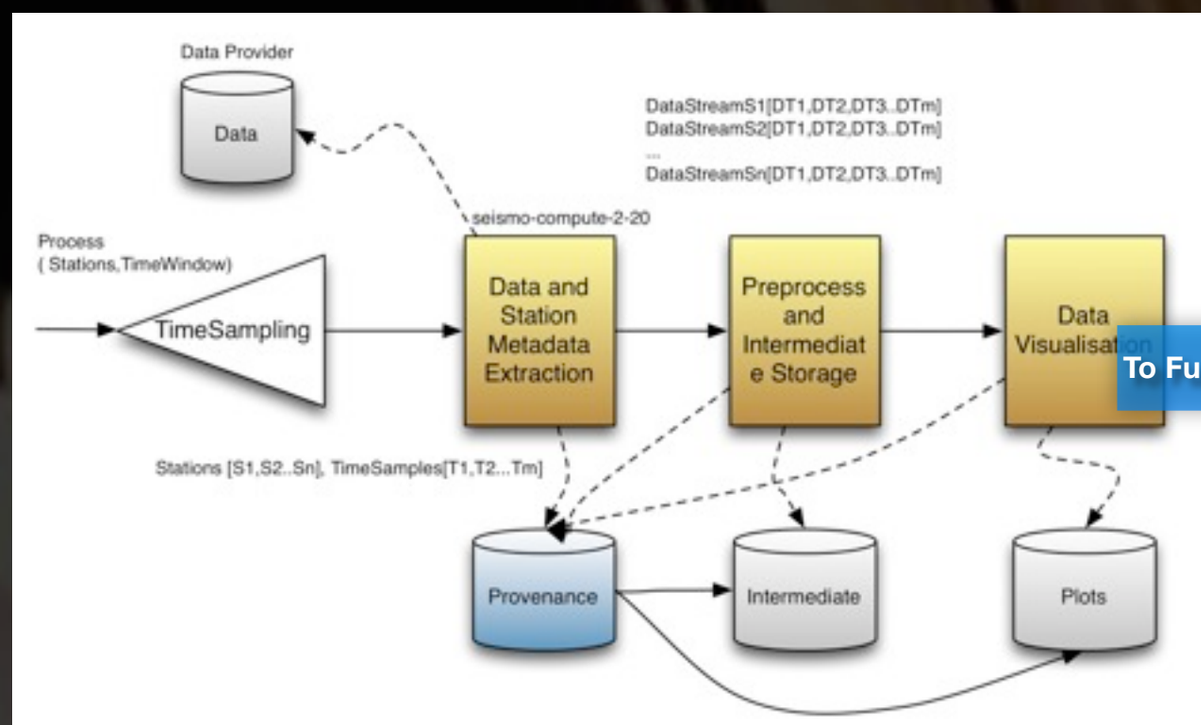
Cross Correlation
[Zaccarelli et al. 2011]

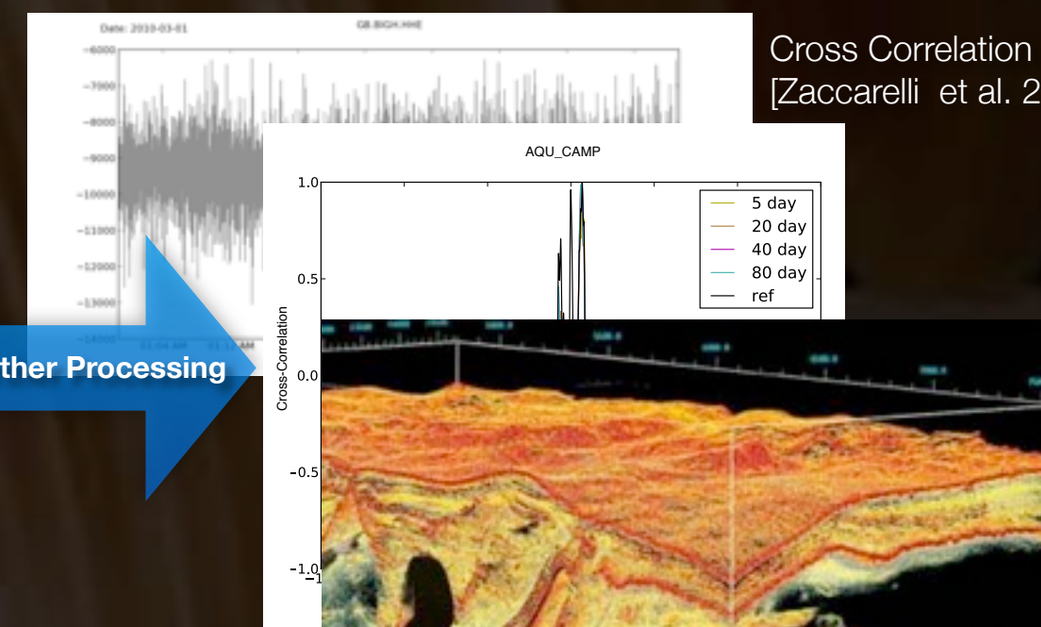
Seismological Pipelines

Processing Chains are common practice in

Seismic Interferometry [G. D. Bensen et al. 2007]

Forward Modelling [A. Fichter et al. 2008]





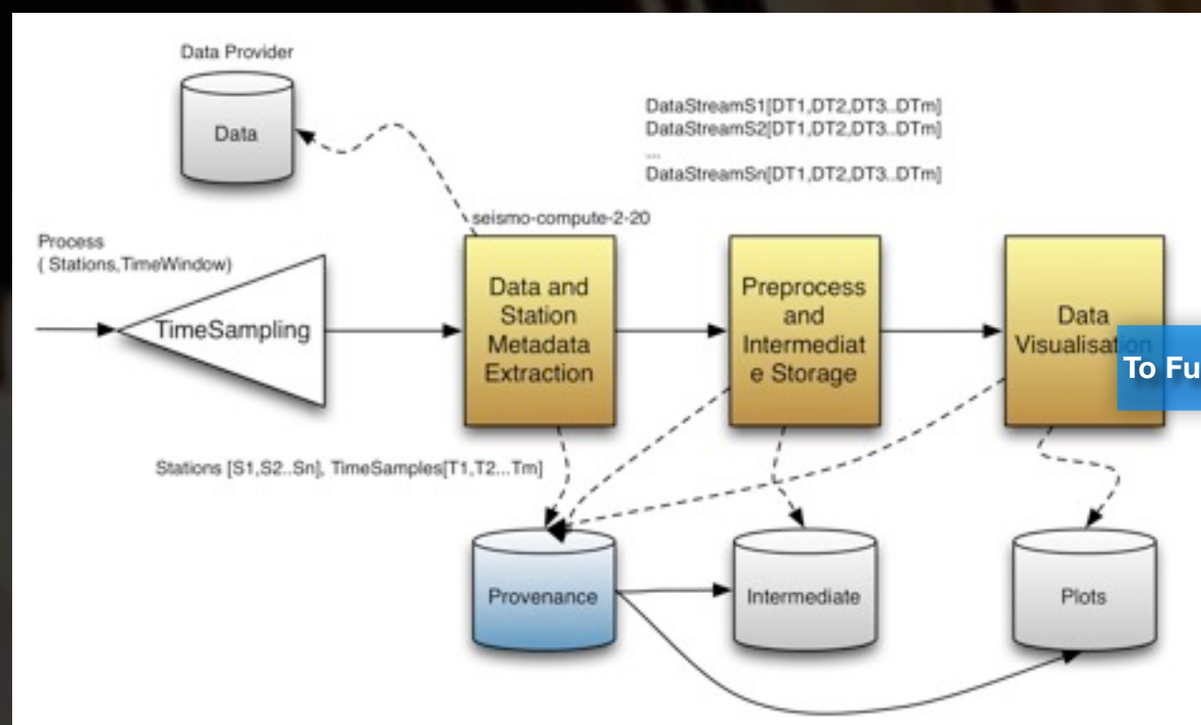
Cross Correlation
[Zaccarelli et al. 2011]

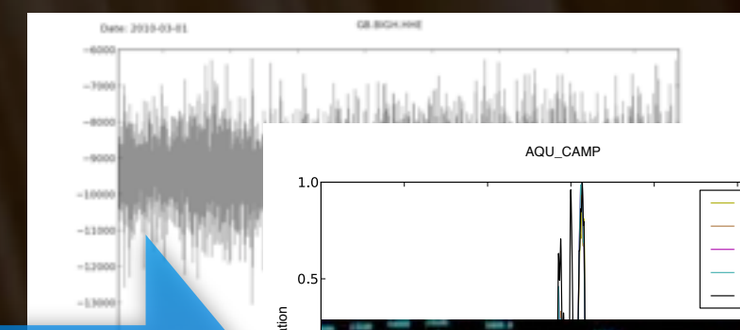
Seismological Pipelines

Processing Chains are common practice in

Seismic Interferometry [G. D. Bensen et al. 2007]

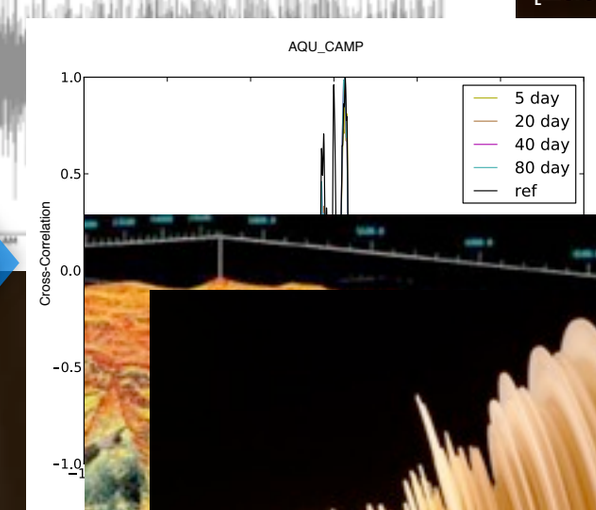
Forward Modelling [A. Fichter et al. 2008]





Cross Correlation
[Zaccarelli et al. 2011]

To Further Processing



Computational Requirements

Seismic Interferometry:

Data Intensive: High IO/CPU rate.

Data processing distribution is the main concern.

Distributed Streaming model suitable for in-memory and time-series data analysis.

Forward Modeling:

Execution of established simulation software (Seissol, Specfem3D) on HPC resources

Large number of synthetics waveforms (time-series) produced can be **streamed** and processed on Data Intensive Systems

Scientific Workflows:

Unified framework for the execution and validation of scientific applications over scalable and distributed resources

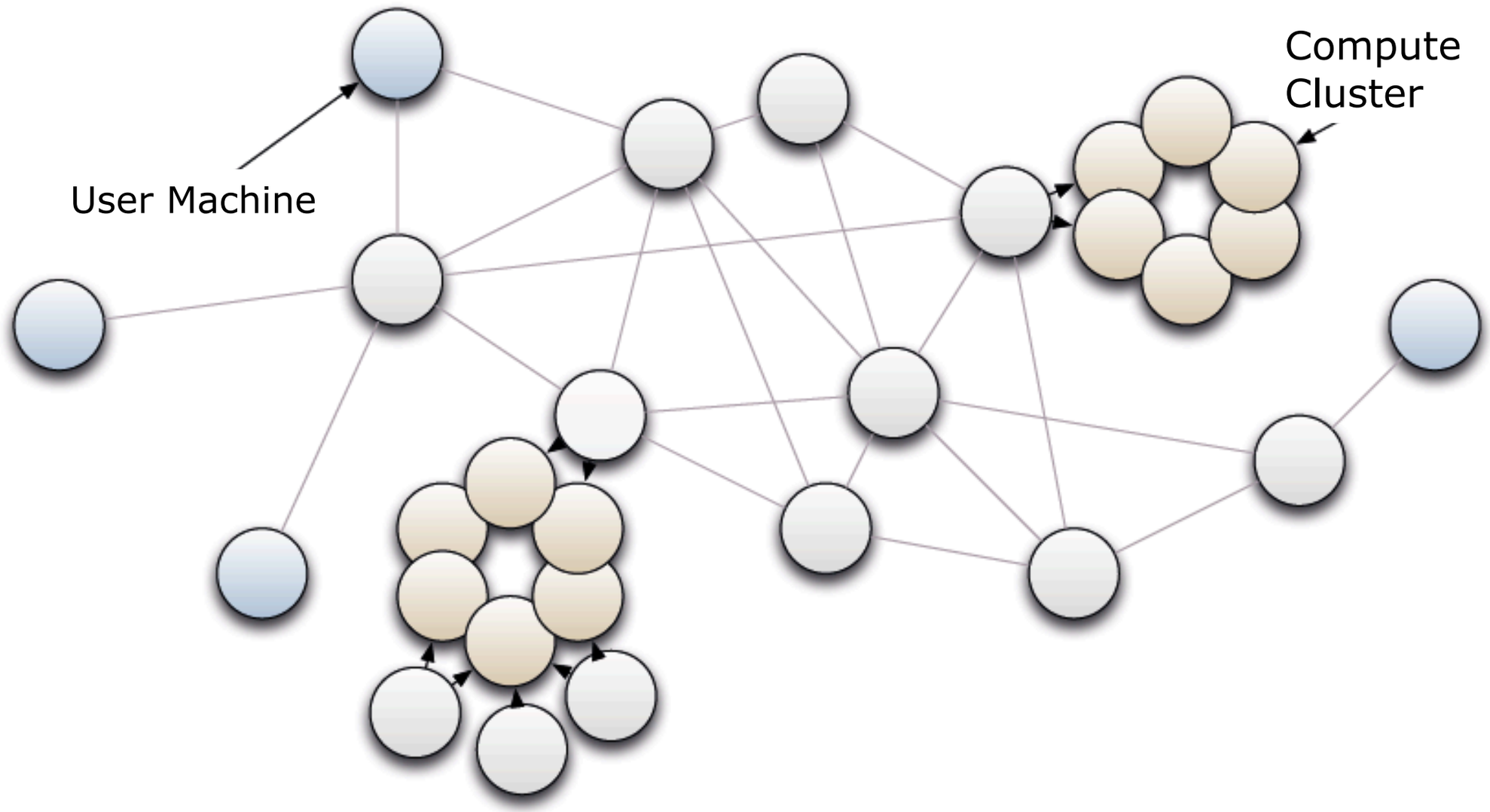
DISPEL (WF Language and Architecture):

- Abstraction on the Data Management Layer (Retrieval, Intermediate Results)
- Abstraction on the enactment and the Computational resources

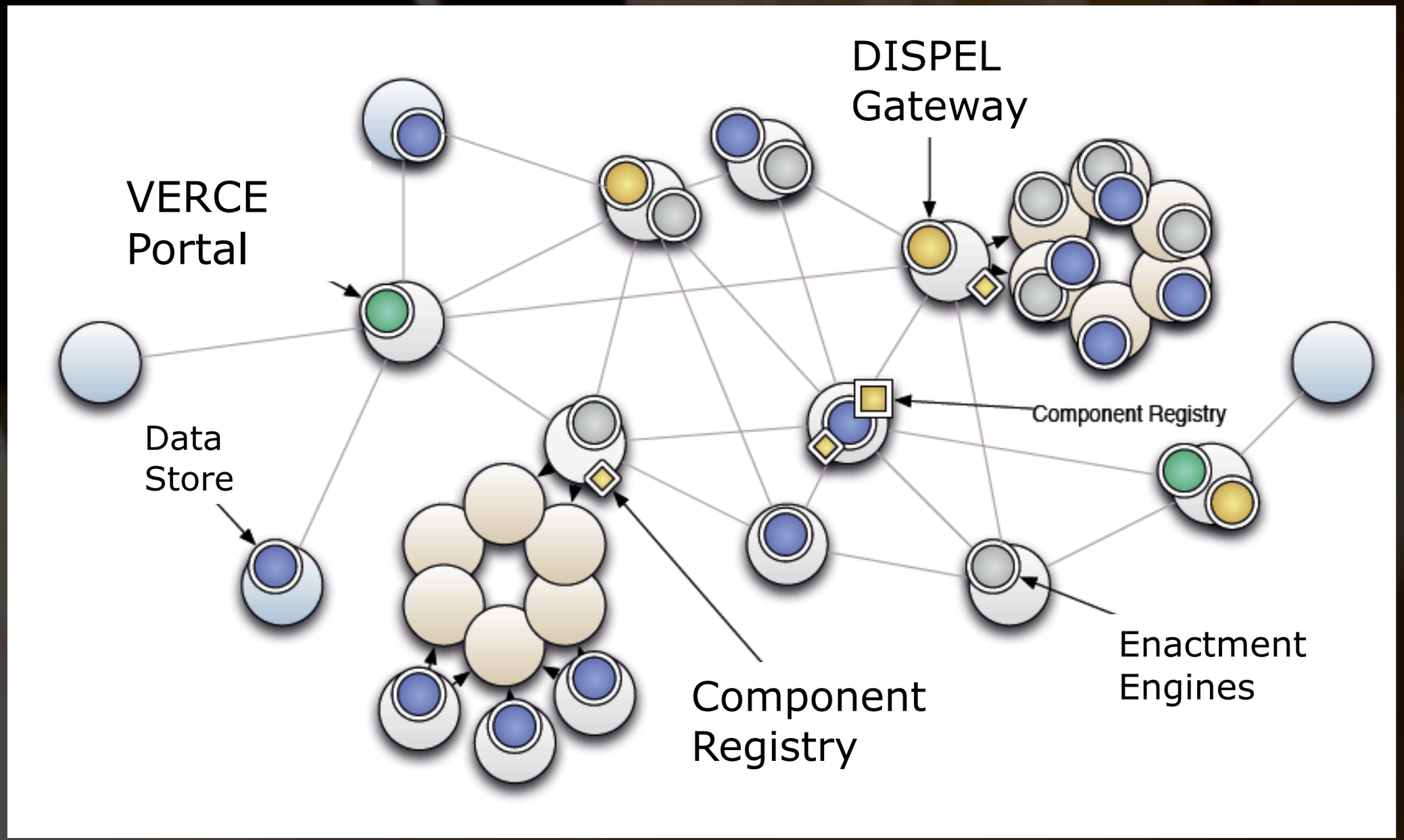
Modular and Monitorable:

- User Analysis Code in Reusable Processing Elements
- Provenance model for Validation and Monitoring of the Results

DISPEL: Abstracts on Distributed Components and Resources



DISPEL: Abstracts on Distributed Components and Resources



Integration of Enactment and User Defined Scripts

Reusable Processing Elements

Integration of Enactment and User Defined Scripts

Reusable Processing Elements

User Code Integration framework:

Easy for scientists to implement, hides most of the WFMS complexity.

Multi-platform and command line testing (JSON I/O protocol).

```
$ cat < inputfile > | python <yourscript> <VERCE JSON> <PARAMETERS JSON (defined
accordinglyto the PE parameters)> > <outputfile>
```

Automated and extensible metadata extraction (as provenance artifacts):

Artifact's Metadata in JSON

Support for automatic annotations

Integration of Enactment and User Defined Scripts

Reusable Processing Elements

Name	Author
Synchronization	INGV
DataTrimming	LMU
SplitStream	INGV
FillGaps	INGV
TemporalNormalization	LMU
ResamplingINGV	INGV
ResamplingWHISPER	CNRS
Whitening	INGV
InstrumentCorrection	KNMI
Detrend	LMU
Filter	LMU
MapSeissolOutputToStations	LMU
MisfitCalculation	LMU

Name	Author
StreamToSeedFile	KNMI
CorrelateNoiseLMU	LMU
CorrelateNoiseWHISPER	CNRS
CorrelateNoiseINGV	INGV
WaveformPlot	INGV
Spectrogram	LMU
ArrayStacking	LMU
StackPlot	LMU
MonetDBFileStream	KNMI
ConvertSeissolOutput	LMU
Specfem3d2Stream	INGV

VERCE

Roles of Provenance

Provenance can describe the analysis at various granularities (Resources, WF Configuration, Lineage):

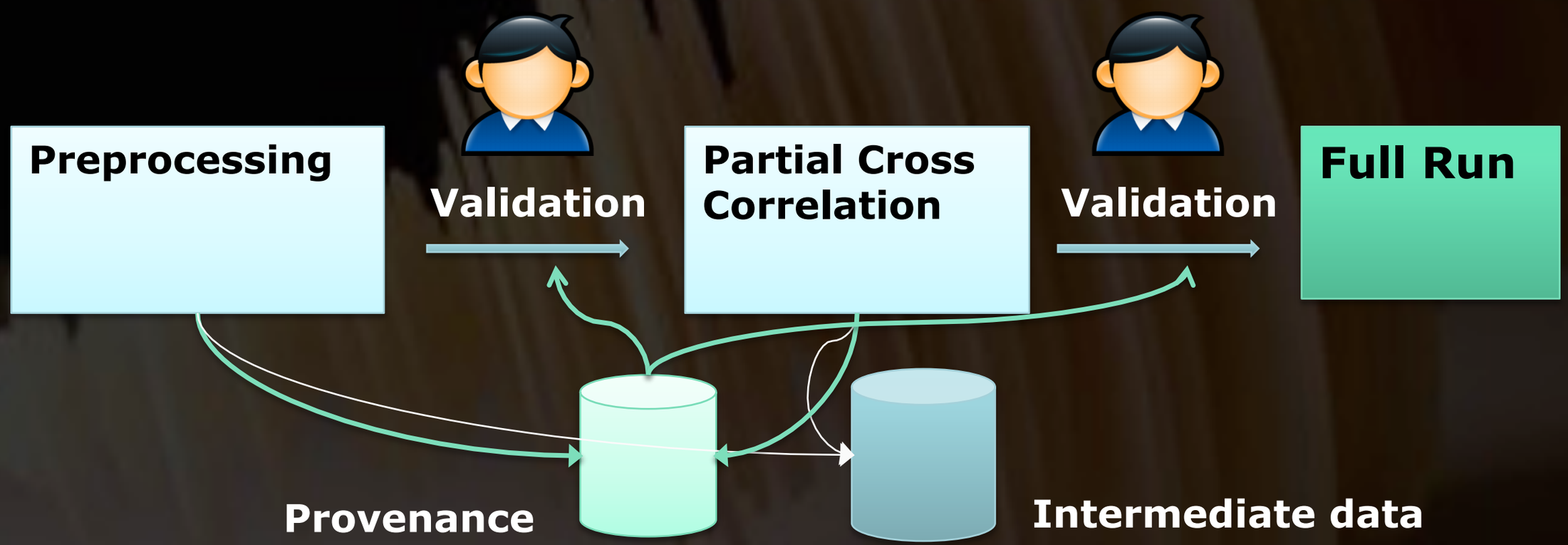
Provide intermediate data visualisation/access for validation at runtime

Process and store heterogeneous/user-defined metadata, processing parameters and errors

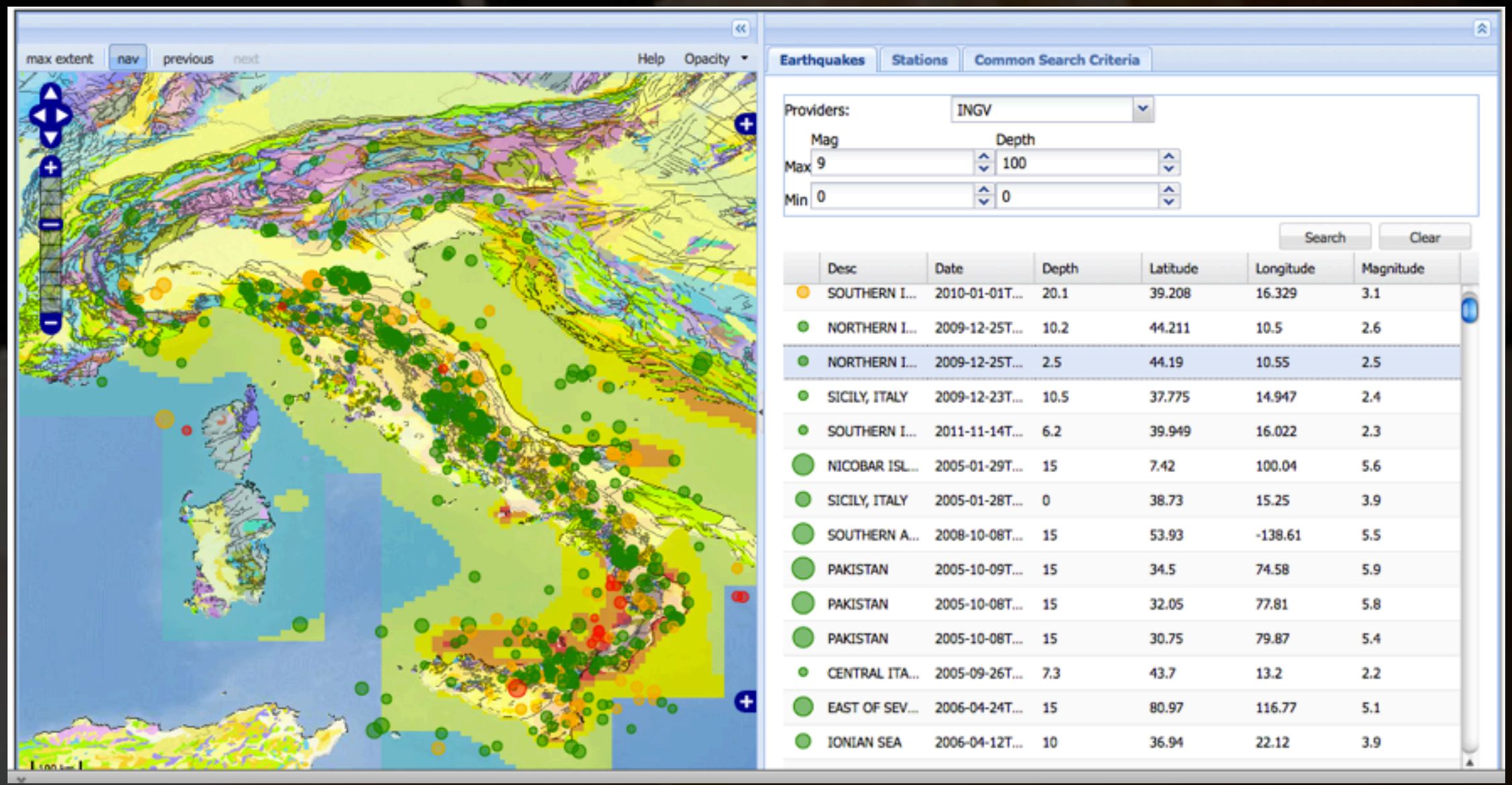
Capture and describe data interdependencies

Aware of Mutations in Software (PE), Machines and Data Sources (Registry)

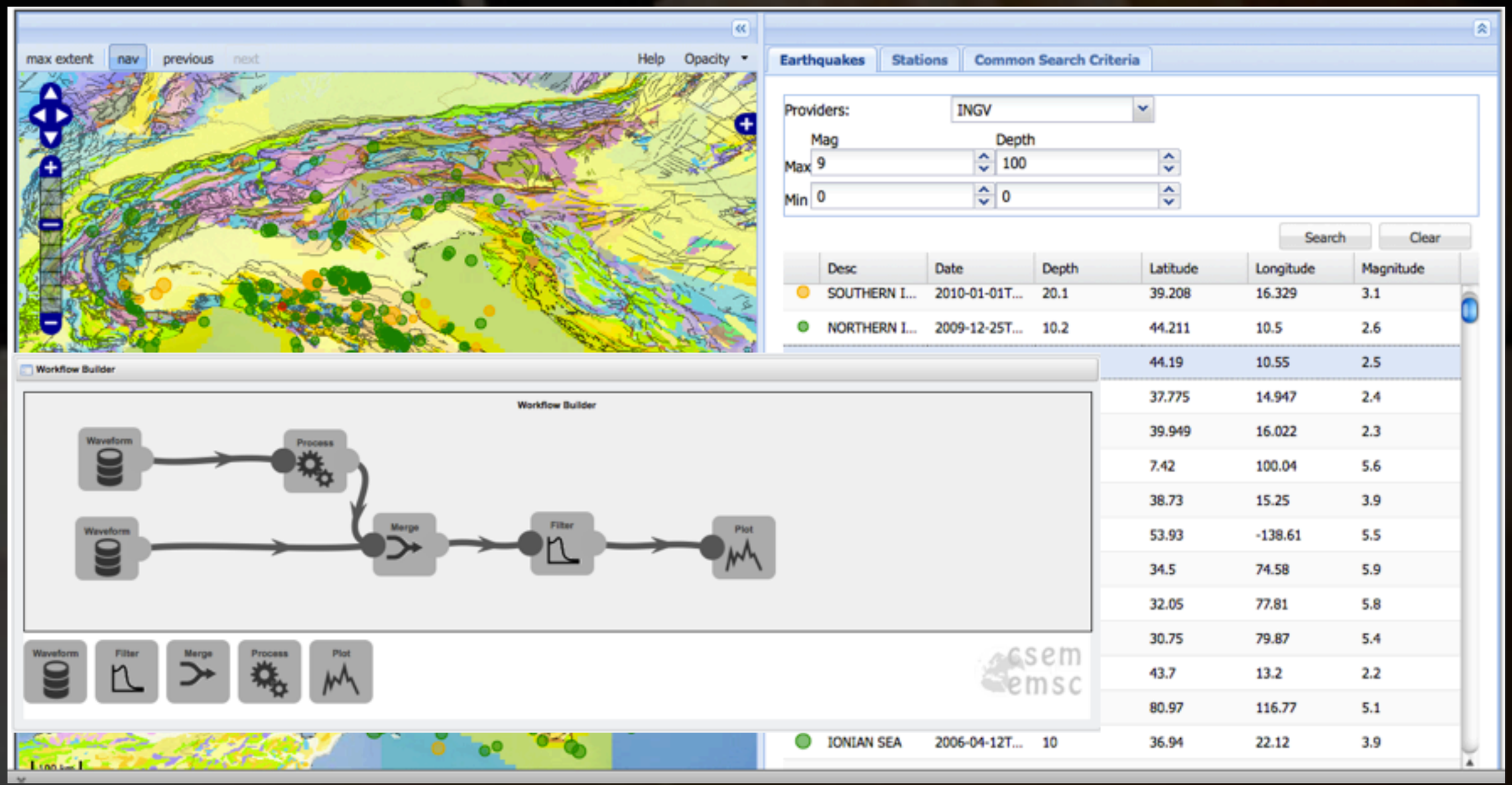
Roles of Provenance



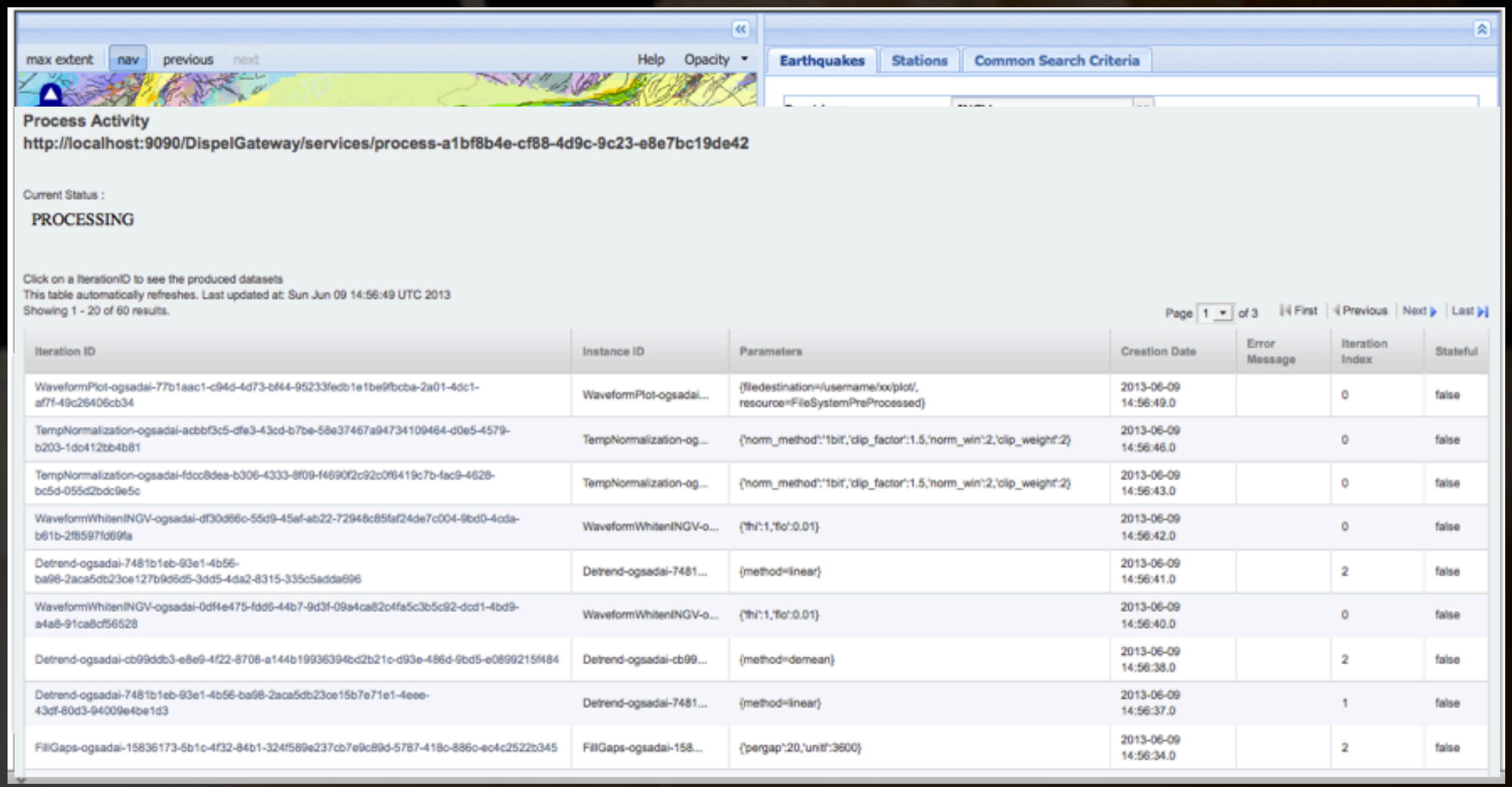
In Progress: Custom access within dedicated Science Gateways accessing external datasources



In Progress: Custom access within dedicated Science Gateways accessing external datasources

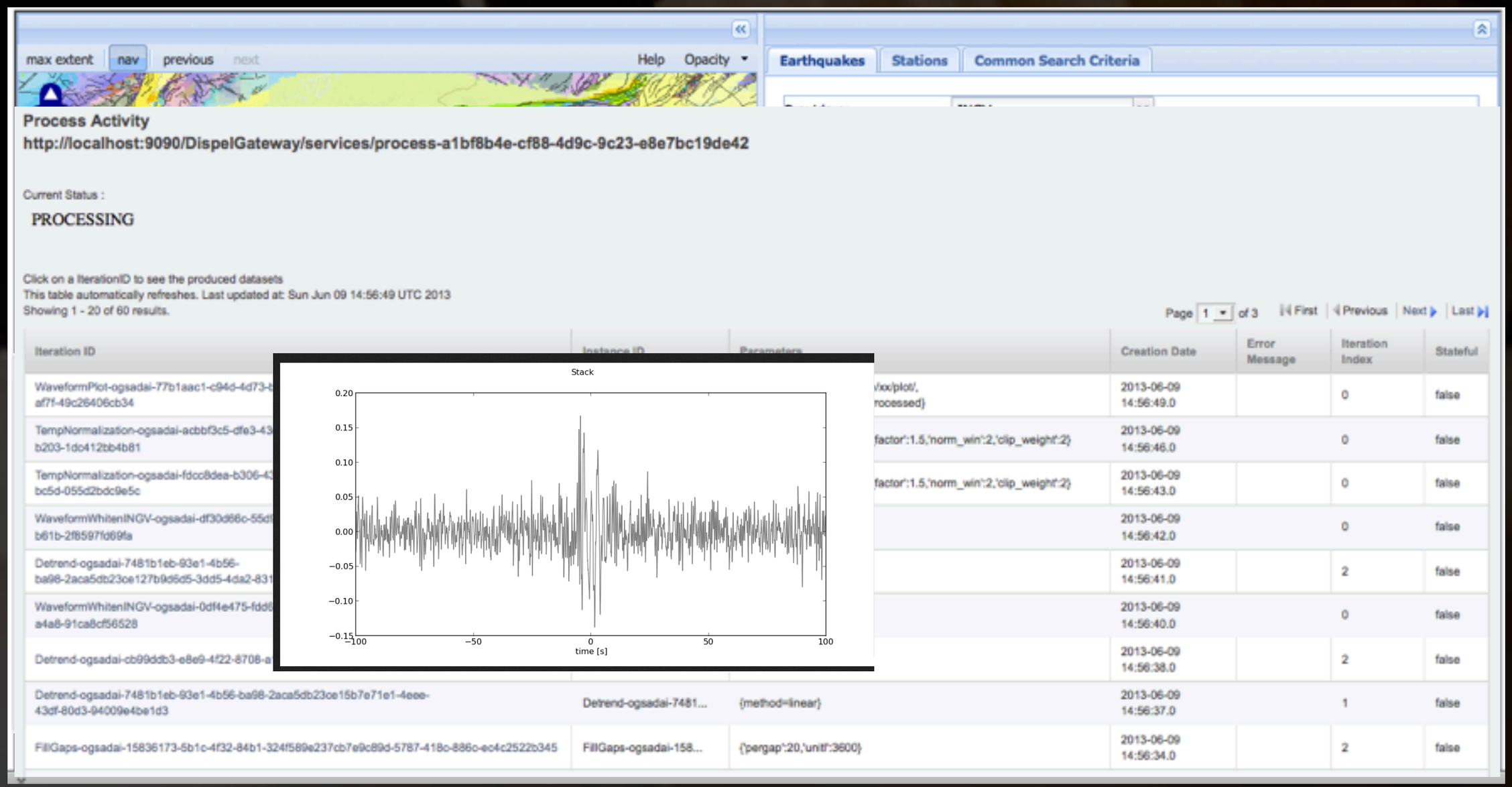


In Progress: Custom access within dedicated Science Gateways accessing external datasources

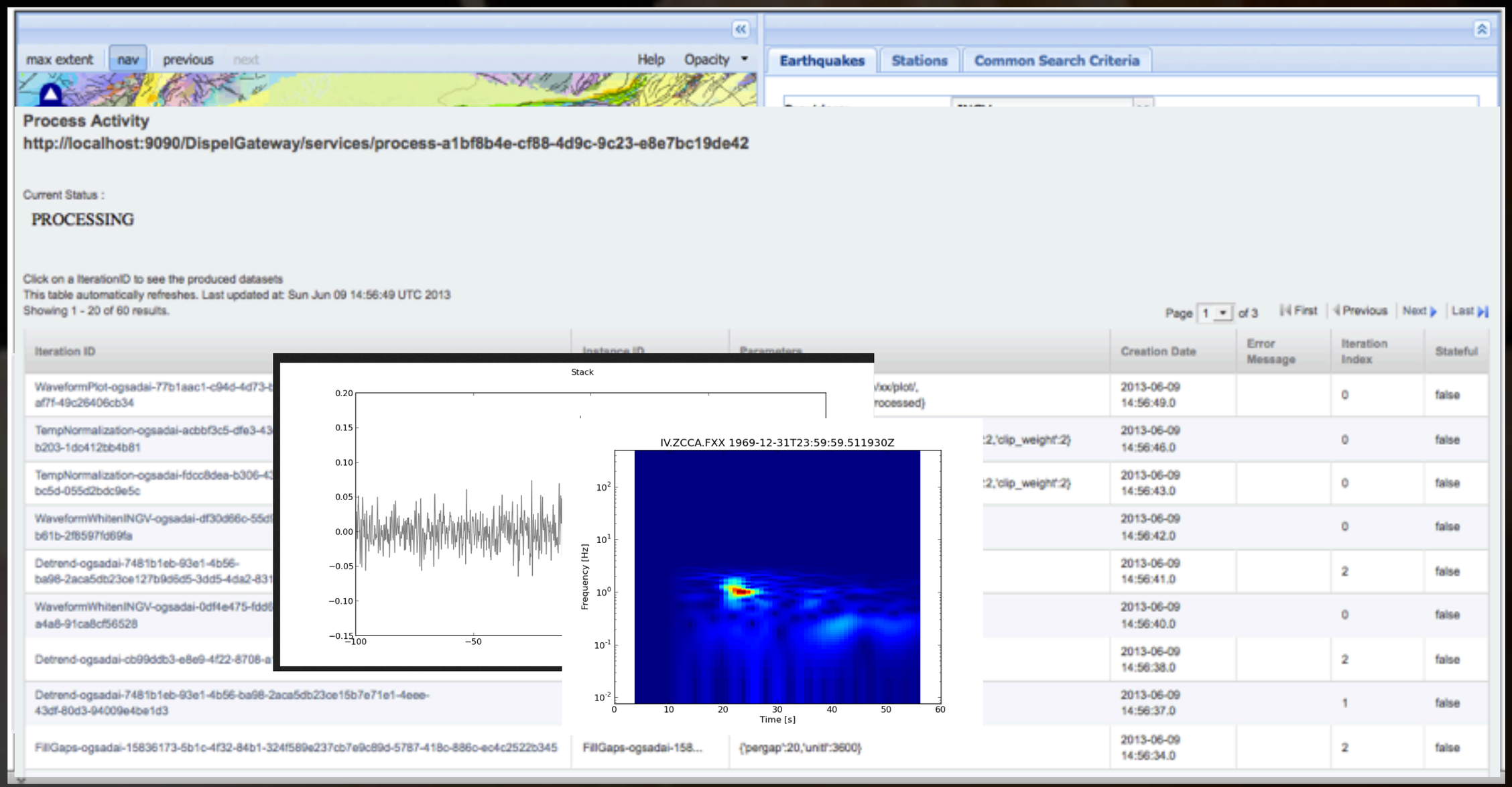


Iteration ID	Instance ID	Parameters	Creation Date	Error Message	Iteration Index	Stateful
WaveformPlot-ogsadai-77b1aac1-c94d-4d73-bf44-95233fedb1e1be9fbcba-2a01-4dc1-af7f-49c26406cb34	WaveformPlot-ogsadai...	{filedestination=/usernamel/xx/plot/, resource=FileSystemPreProcessed}	2013-06-09 14:56:49.0		0	false
TempNormalization-ogsadai-acbbf3c5-dfe3-43cd-b7be-58e37467a94734109464-d0e5-4579-b203-1dc412bb4b81	TempNormalization-og...	{'norm_method':'1bit','clip_factor':1.5,'norm_win':2,'clip_weight':2}	2013-06-09 14:56:46.0		0	false
TempNormalization-ogsadai-fdcc8dea-b306-4333-8f09-f4690f2c92c0f6419c7b-fac9-4628-bc5d-055d2bdc9e5c	TempNormalization-og...	{'norm_method':'1bit','clip_factor':1.5,'norm_win':2,'clip_weight':2}	2013-06-09 14:56:43.0		0	false
WaveformWhitenINGV-ogsadai-df30d66c-55d9-45af-ab22-72948c85faf24de7c004-9bd0-4cda-b61b-2f8597fd69fa	WaveformWhitenINGV-o...	{'fhi':1,'flo':0.01}	2013-06-09 14:56:42.0		0	false
Detrend-ogsadai-7481b1eb-93e1-4b56-ba98-2aca5db23ce127b9d6d5-3dd5-4da2-8315-335c5adda696	Detrend-ogsadai-7481...	{method=linear}	2013-06-09 14:56:41.0		2	false
WaveformWhitenINGV-ogsadai-0df4e475-fdd6-44b7-9d3f-09a4ca82c4fa5c3b5c92-dcd1-4bd9-a4a8-91ca8cf56528	WaveformWhitenINGV-o...	{'fhi':1,'flo':0.01}	2013-06-09 14:56:40.0		0	false
Detrend-ogsadai-cb99ddb3-e8e9-4f22-8708-a144b19936394bd2b21c-d93e-486d-9bd5-e0899215f484	Detrend-ogsadai-cb99...	{method=demean}	2013-06-09 14:56:38.0		2	false
Detrend-ogsadai-7481b1eb-93e1-4b56-ba98-2aca5db23ce15b7e71e1-4eee-43df-80d3-94009e4be1d3	Detrend-ogsadai-7481...	{method=linear}	2013-06-09 14:56:37.0		1	false
FillGaps-ogsadai-15836173-5b1c-4f32-84b1-324f589e237cb7e9c89d-5787-418c-886c-ec4c2522b345	FillGaps-ogsadai-158...	{'pergap':20,'unitf':3600}	2013-06-09 14:56:34.0		2	false

In Progress: Custom access within dedicated Science Gateways accessing external datasources



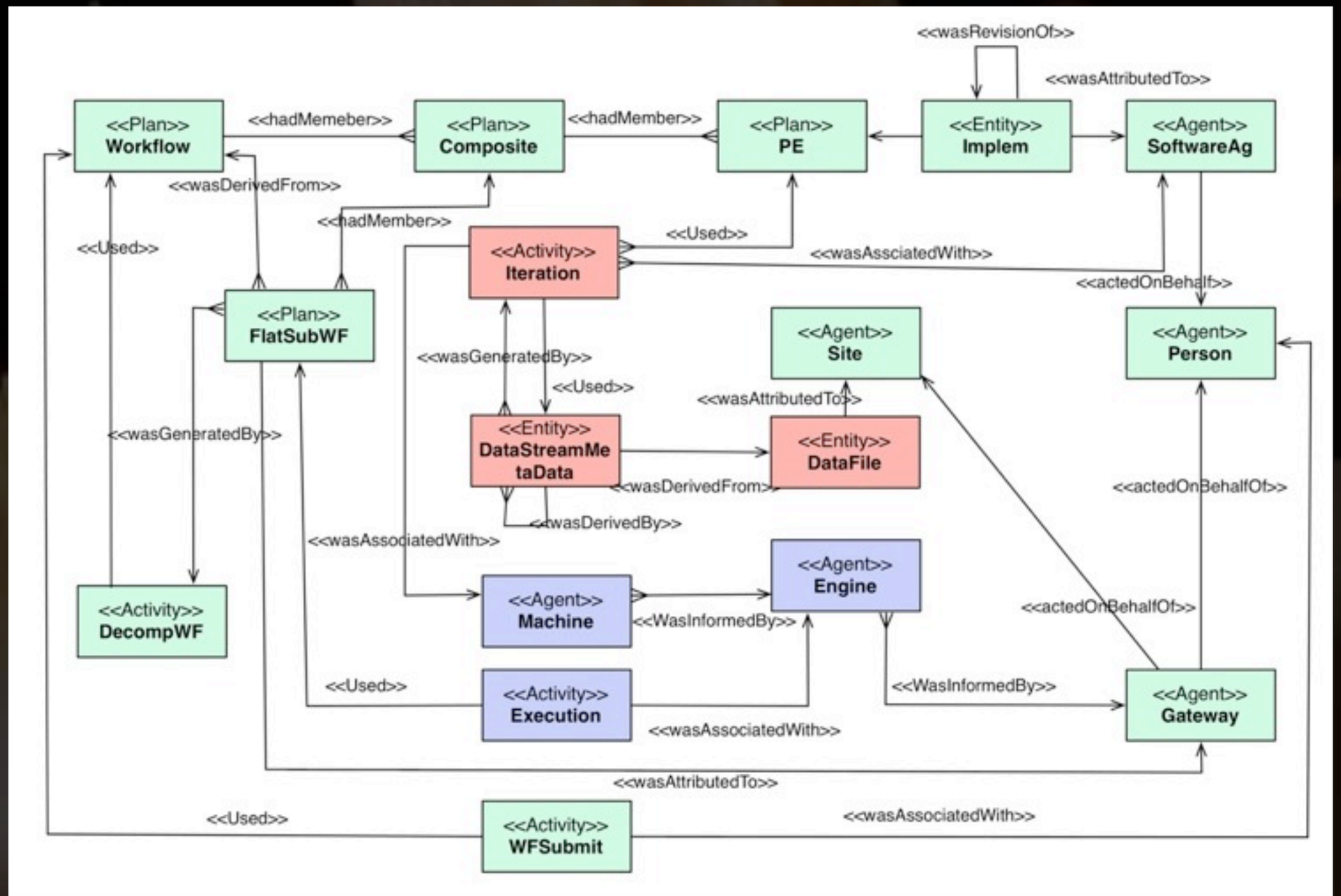
In Progress: Custom access within dedicated Science Gateways accessing external datasources



Compound Model based on W3C PROV

Concepts Distribution

- Lineage / Retrospective Prov
- DISPEL/Registry Prospective Prov
- Enactment / Environmental Optimisation



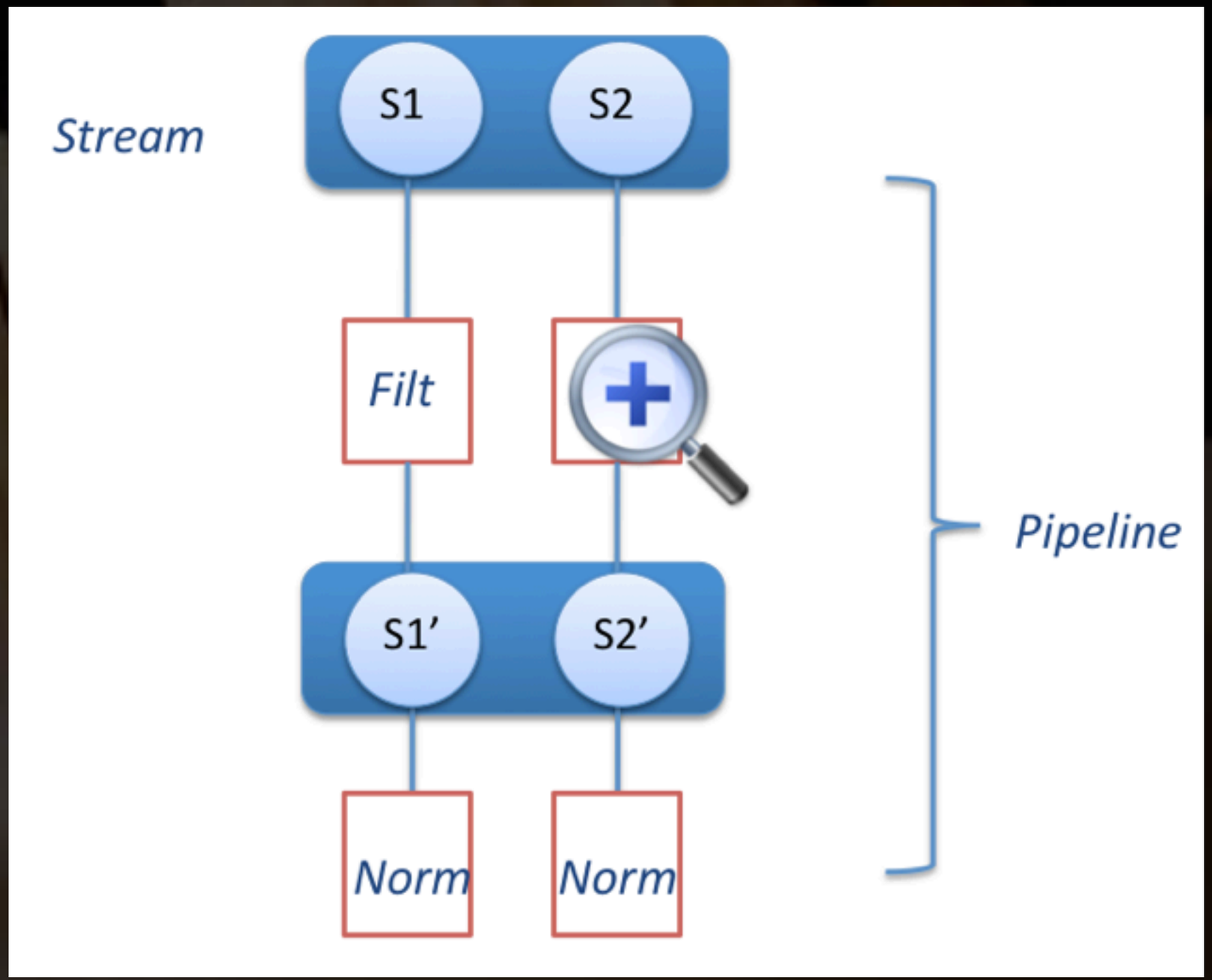
Usage Scenarios

1 - Components Behaviours: Internal transformations within a group of PEs.

2 - Data transformation snap-shots: Runtime view on the variation of a portion of the data stream

3 - Comparing Runs: Comparing 1 & 2 over multiple runs

4 - Smart Reruns and Fault-tolerance: Checkpoints could trigger the storage of the data



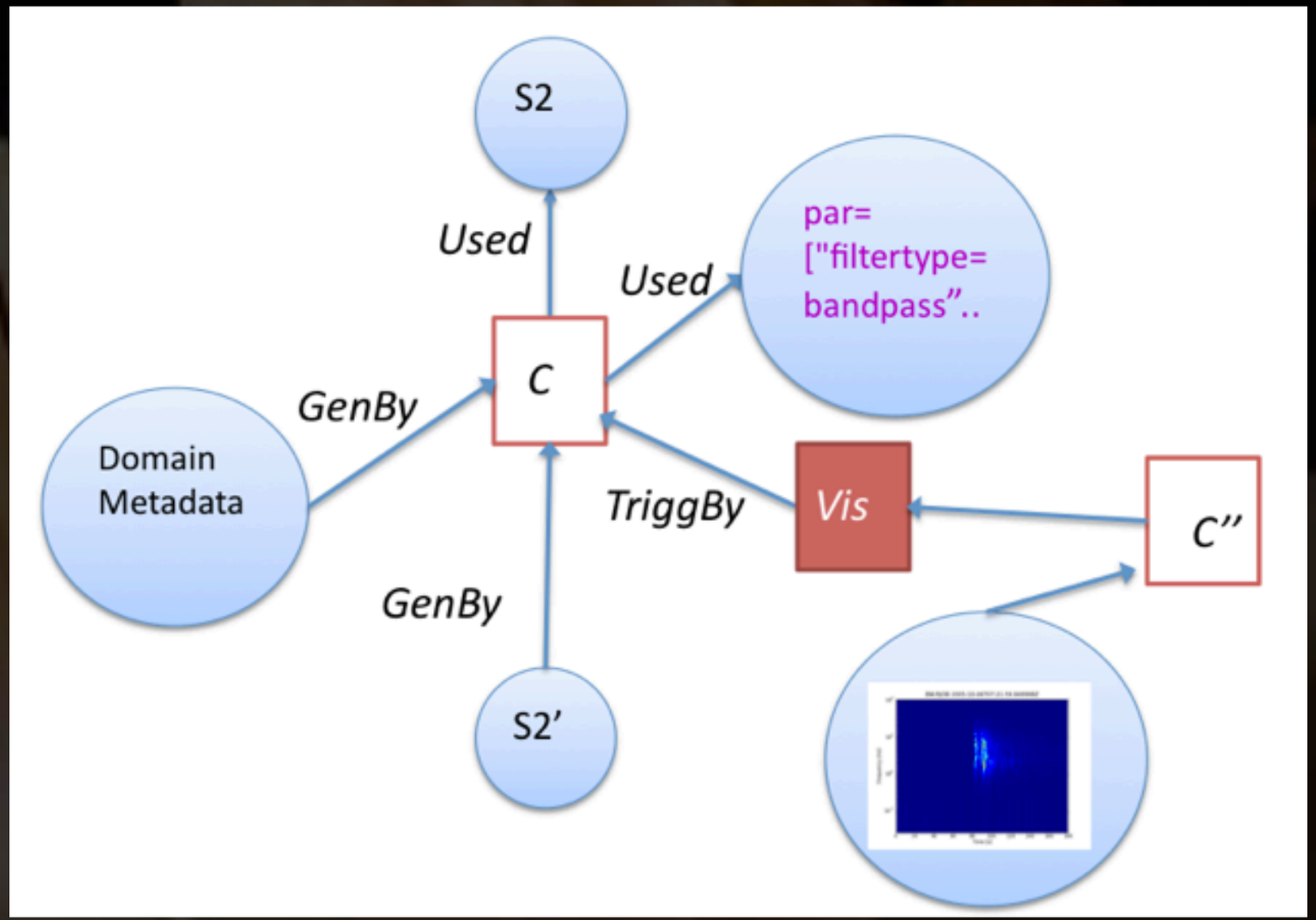
Usage Scenarios

1 - Components Behaviours: Internal transformations within a group of PEs.

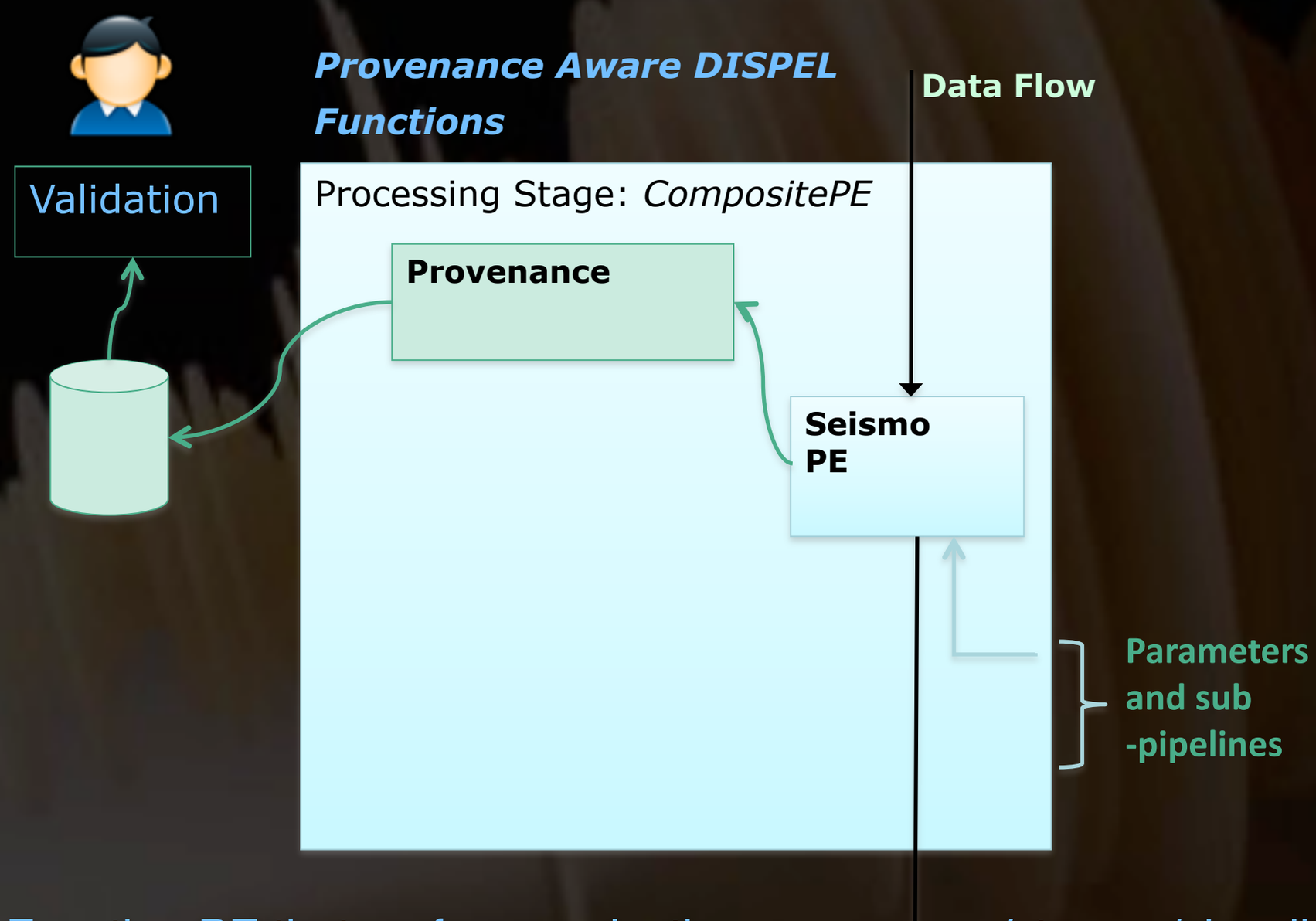
2 - Data transformation snap-shots: Runtime view on the variation of a portion of the data stream

3 - Comparing Runs: Comparing 1 & 2 over multiple runs

4 - Smart Reruns and Fault-tolerance: Checkpoints could trigger the storage of the data

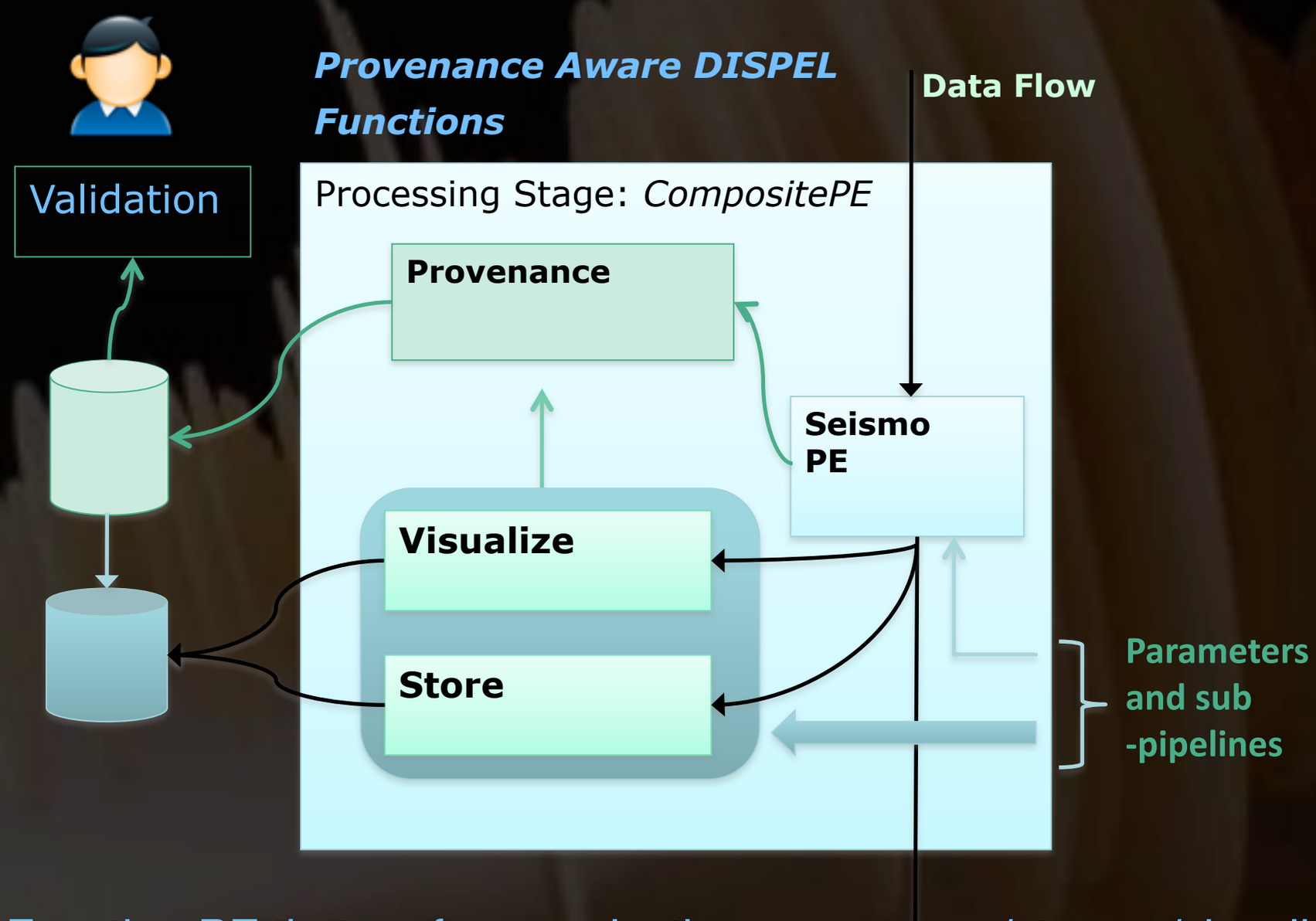


DISPEL functions and Pipeline's <*SeismoStages*>



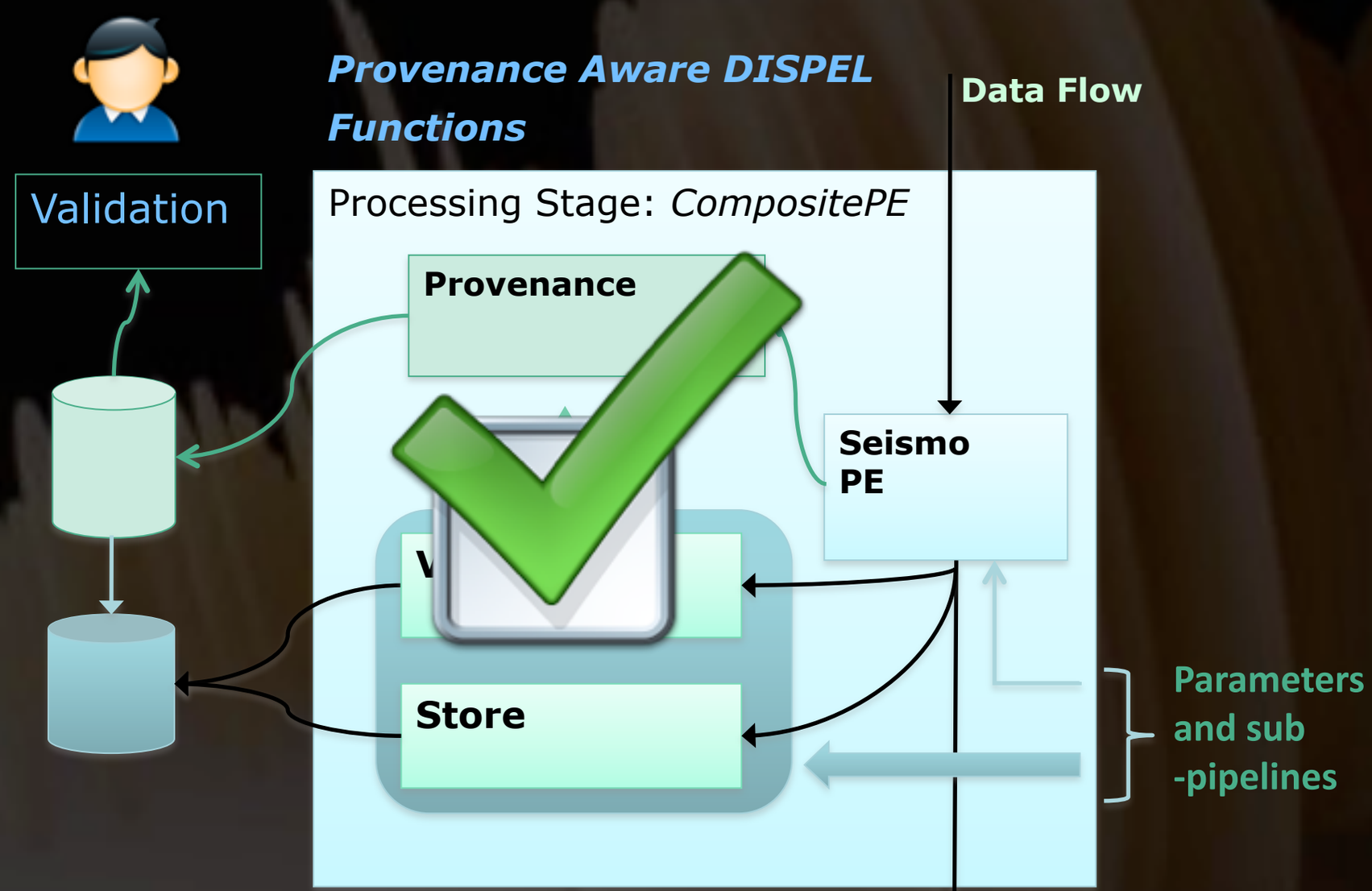
DISPEL Function PE that performs selective provenance/storage/visualisation of the processed datastream

DISPEL functions and Pipeline's <*SeismoStages*>



DISPEL Function PE that performs selective provenance/storage/visualisation of the processed datastream

DISPEL functions and Pipeline's <*SeismoStages*>



DISPEL Function PE that performs selective provenance/storage/visualisation of the processed datastream

Harvesting Challenges

- Preparatory phases: The evaluation of various processing strategies requires multiple runs adopting different components and configurations
- Full coverage of retrospective provenance could present the following numbers:

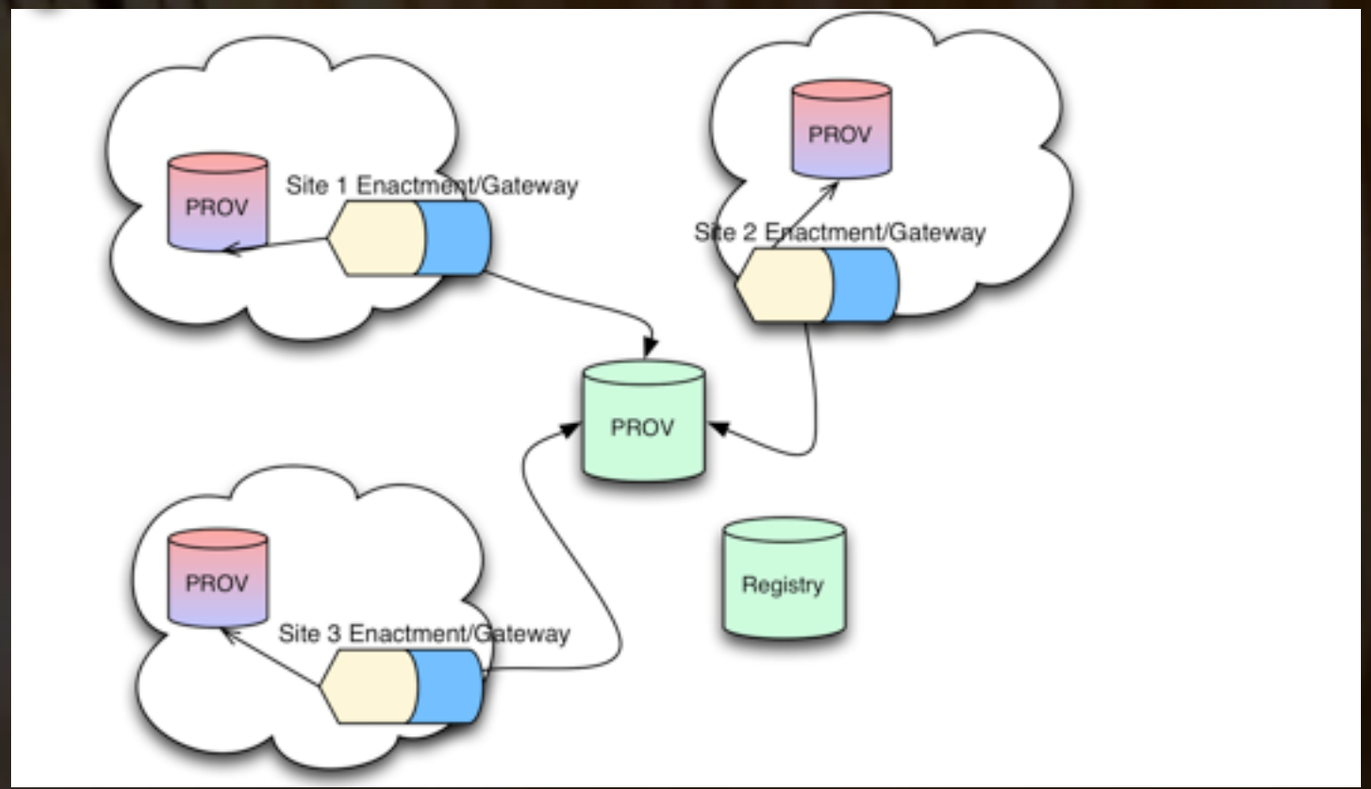
  Cross Correlation, preparatory phase, single run:

  - Preprocessing: 20 stations, one year of data, 1 hour chunks size, 6 preprocessing stages = 1.051.200 artifacts
  - Cross correlation: (20(20-1))/2 x-correlation pairs plus data stacking = 3.328.800 artifacts.
  - **We are looking at a selective recoding strategies.** Decrease I/O rate and the storage of not interesting metadata

Harvesting Challenges

Centralised vs Distributed Storage:

- Computation can be distributed across different models and institutional infrastructures.
- Computation can happen in locations with no persistent storage facilities, while provenance data has to be preserved



Harvesting Challenges

Different enactment engines

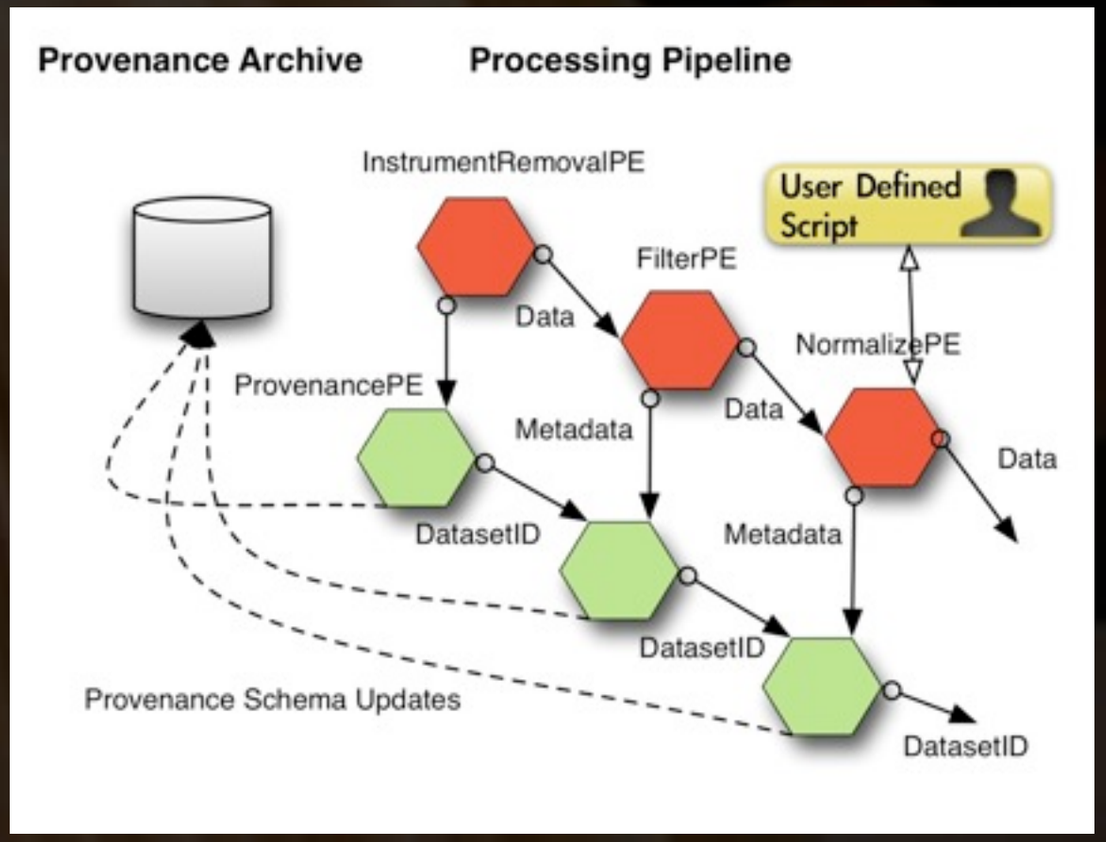
Not all of them support provenance (OGSA-DAI, Twitter STORM, Apache S4)

Strategies in Evaluation:

In WF, programmatic provenance

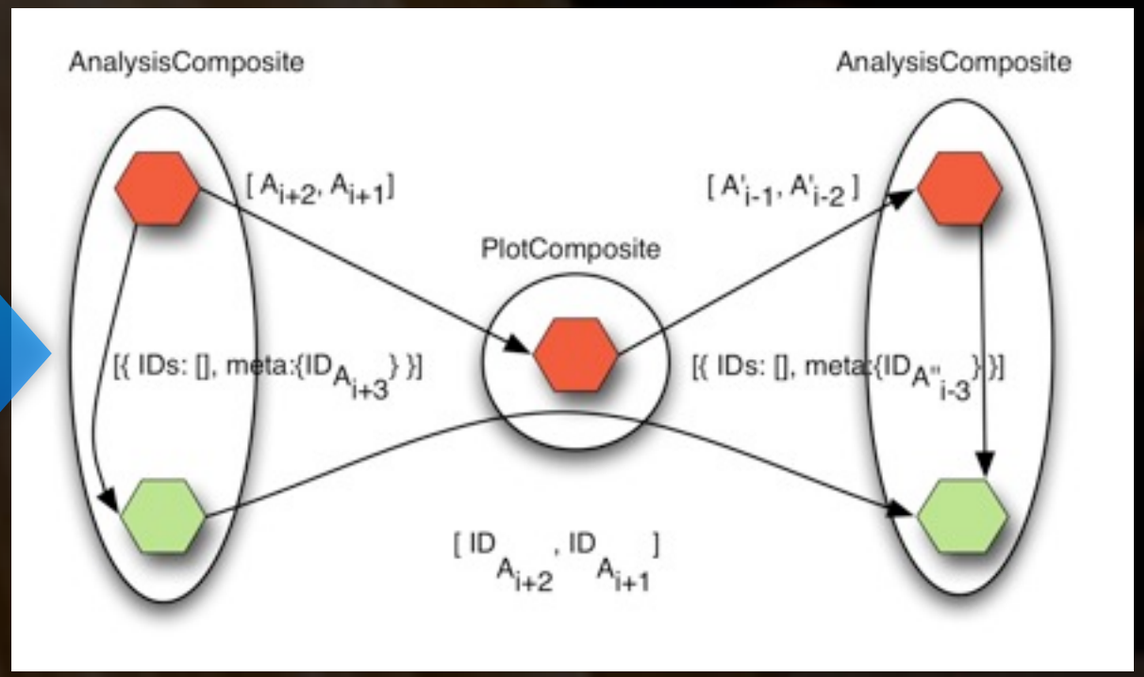
DISPEL functions returning configurable and provenance aware CompositePEs

Provenance PEs vs **semantic annotations on functions**



DISPEL functions and Pipeline's <*SeismoStages*>

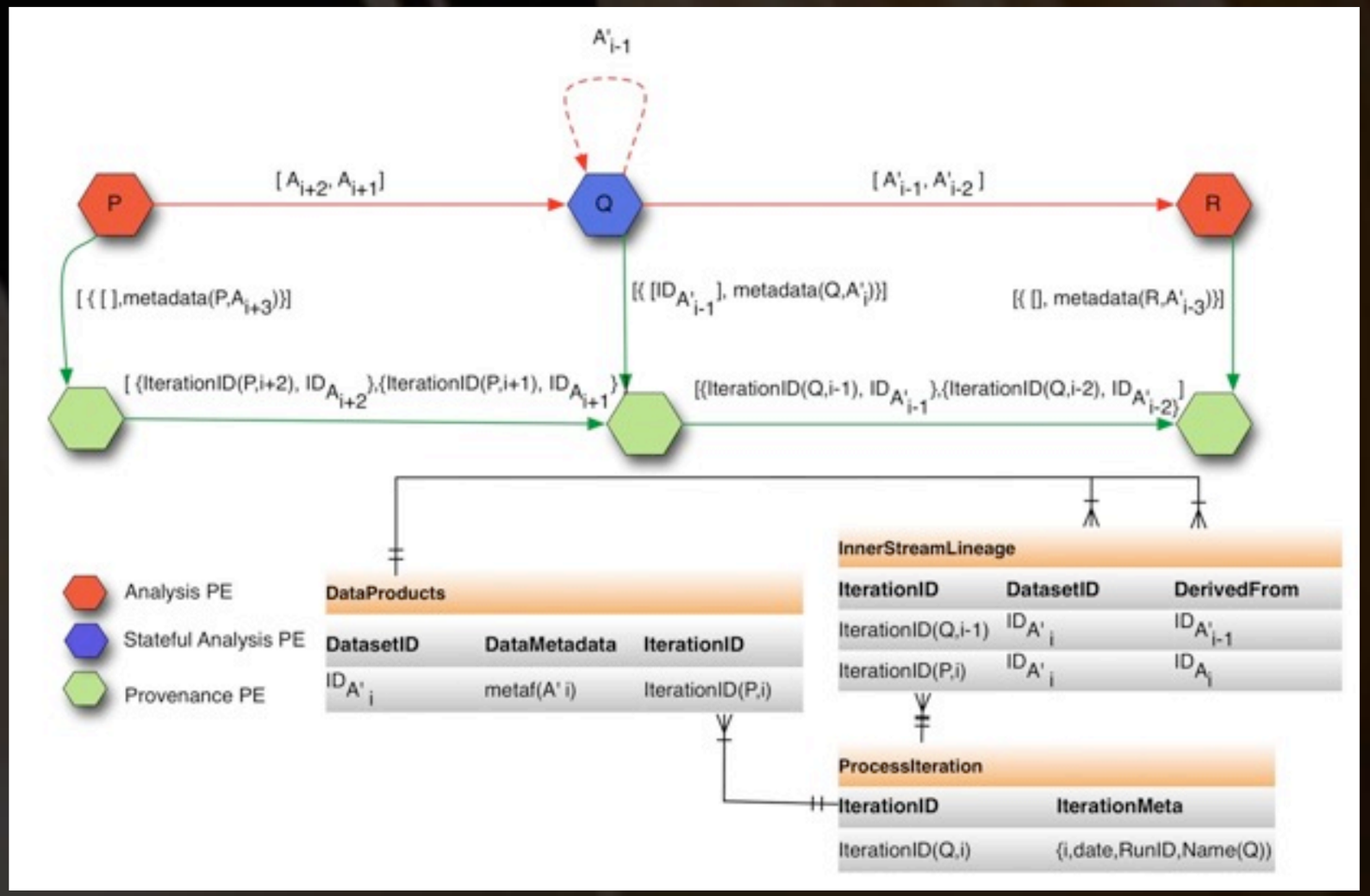
```
PE<SeismoStage> plot(String plotLocation,
                     String provenanceResource,
                     String processingResource,
                     Boolean traceit) {

   /* Echo PEs are used to pass the lineage */
   /* data to the provenance PE if traceit is true */
   /* or to the next Stage if traceit is false */

 Echo echoin = new Echo;
 Echo echoout = new Echo;
 ProvenancePE prov = new ProvenancePE;
 WaveformPlot plot = new WaveformPlot;

 /* Provenance recording enabled */
 if (traceit) {

   |-provenanceResource-| => prov.resource;
   plot.metadata => prov.metastring;
   echoin.output => prov.lineagein;
   prov.lineageout => echoout.input;

 /* Provenance recording disabled */
  } else {

      plot.metadata => discard;
      echoin.output=>echoout.input;

    }

 |-processingResource-| => plot.resource;
 |-repeat enough of ["filepath="+plotLocation]-|
      => plot.parameters;

 /* Returns the composite PE */
 return SeismoStage(

      <Connection input=plot.input;
       Connection lineagein= echoin.input>
      =>
      <Connection lineageout= echoout.output;
       Connection output=plot.output>);

};
```



DISPEL Function returning a <*SeismoStage*> PE that performs the plotting of the data stream.
Provenance can be disabled by making *traceit = False*.

Provenance Model for Streaming Systems

Reference Model: Multi Invocation Model [M. K. Anand et al. 2009]
Keeping the Model implementation agnostic to the WFS internal codebase (general specs)



Stateful scenario: Capturing and storing relevant state and data dependencies for a stateful process **Q** at step *i*.

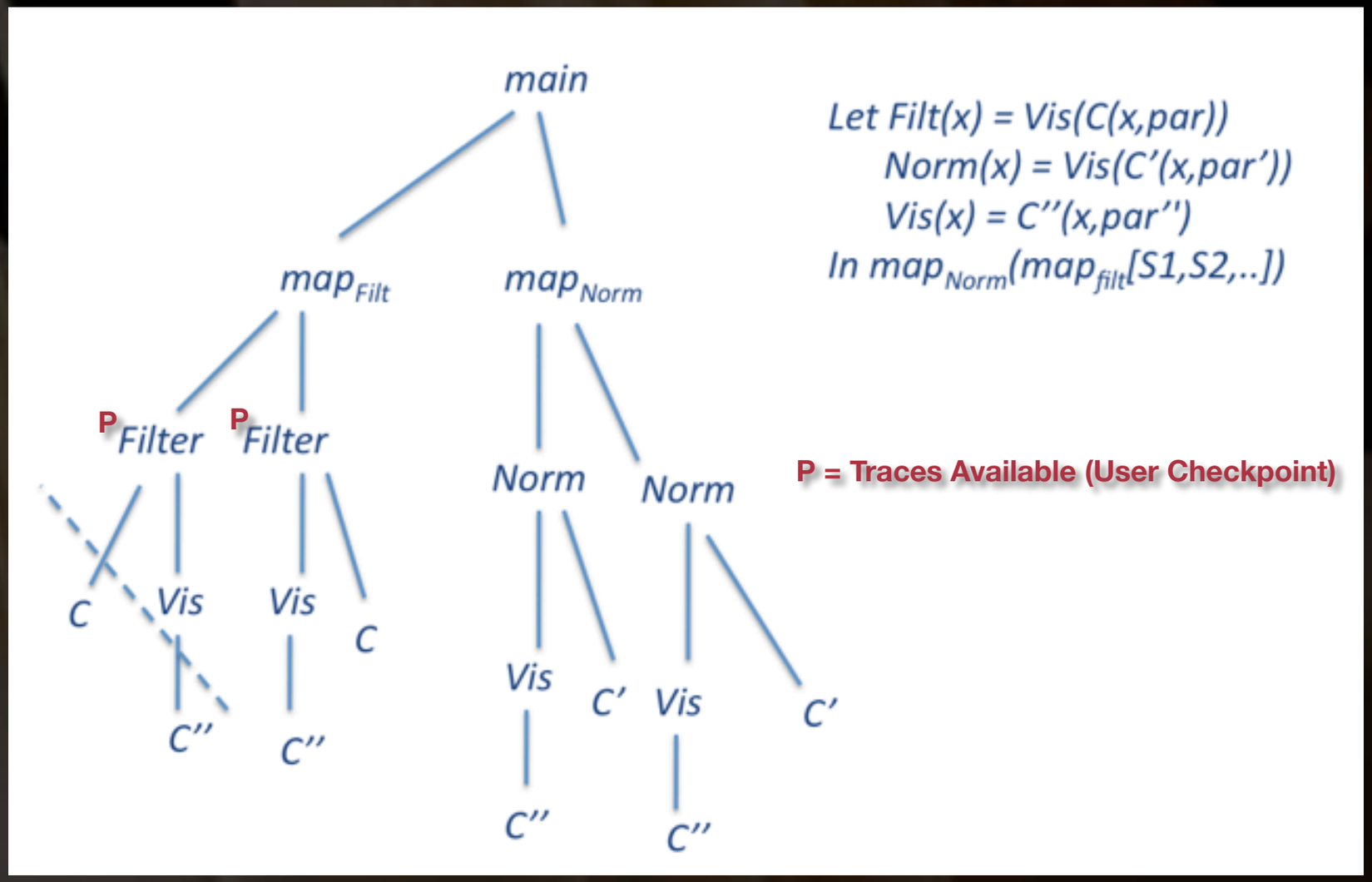
In progress: Functional representation and lineage access via HOPM [Buneman, Cheney, Kostylev, 2012]

Introduction of HOPM like approach (*Hierarchical Model of Provenance*) for DISPEL

Nested hierarchy of function calls.

Expandable where the function is offering provenance recordings (**P**)

Users can define **Checkpoints** to activate provenance harvesting.



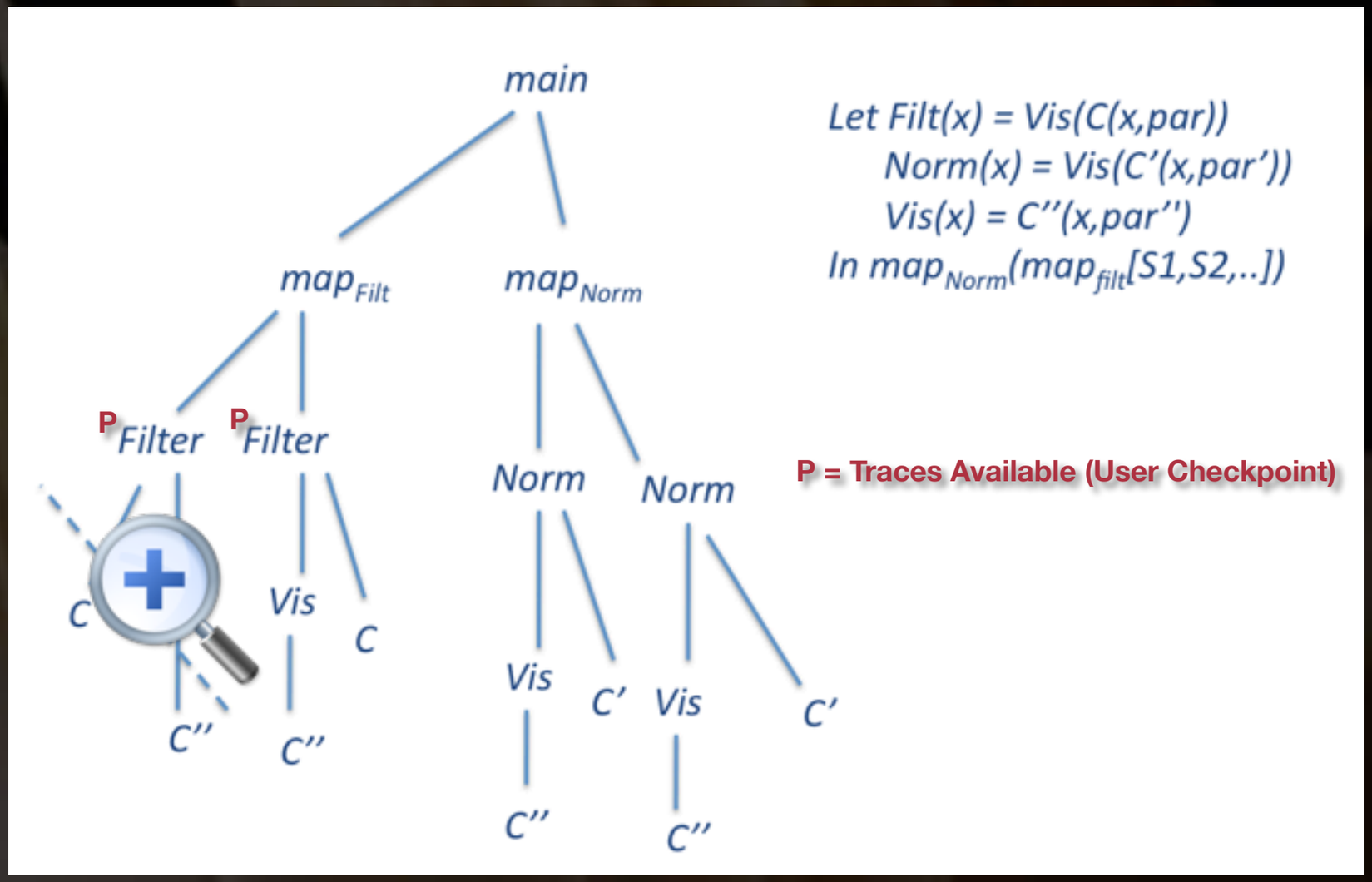
In progress: Functional representation and lineage access via HOPM [Buneman, Cheney, Kostylev, 2012]

Introduction of HOPM like approach (*Hierarchical Model of Provenance*) for DISPEL

Nested hierarchy of function calls.

Expandable where the function is offering provenance recordings (**P**)

Users can define **Checkpoints** to activate provenance harvesting.



Thank you!

VERGE