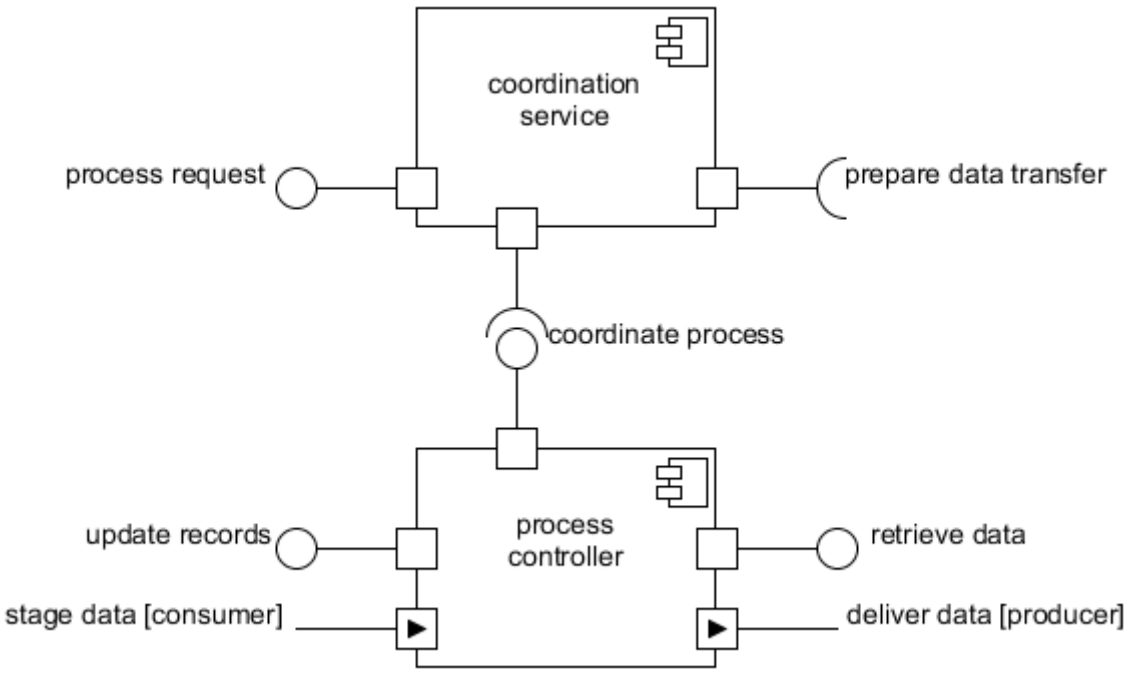
CV Data Processing

The processing of data can be tightly integrated into data handling systems, or can be delegated to a separate set of services invoked on demand. In general, the more complicated processing tasks will require the use of separated services. The provision of dedicated processing services becomes significantly more important when large quantities of data are being curated within a research infrastructure. Scientific data is an example which is often subject to extensive post-processing and analysis in order to extract new results. The data processing objects of an infrastructure encapsulate the dedicated processing services made available to that infrastructure, either within the infrastructure itself or delegated to a client infrastructure.

Data Processing Objects



Notation

CV data processing objects are described as a set of **process controller** (representing the computational functionality of registered execution resources) monitored and managed by a **coordination service**. The coordination service delegates all processing tasks sent to particular execution resources, coordinates multi-stage workflows and initiates execution. Data may need to be **staged** onto individual execution resources and results **persisted** for future use; data channels can be established with resources via their process controllers. The following diagrams shows the staging and persistence of data.

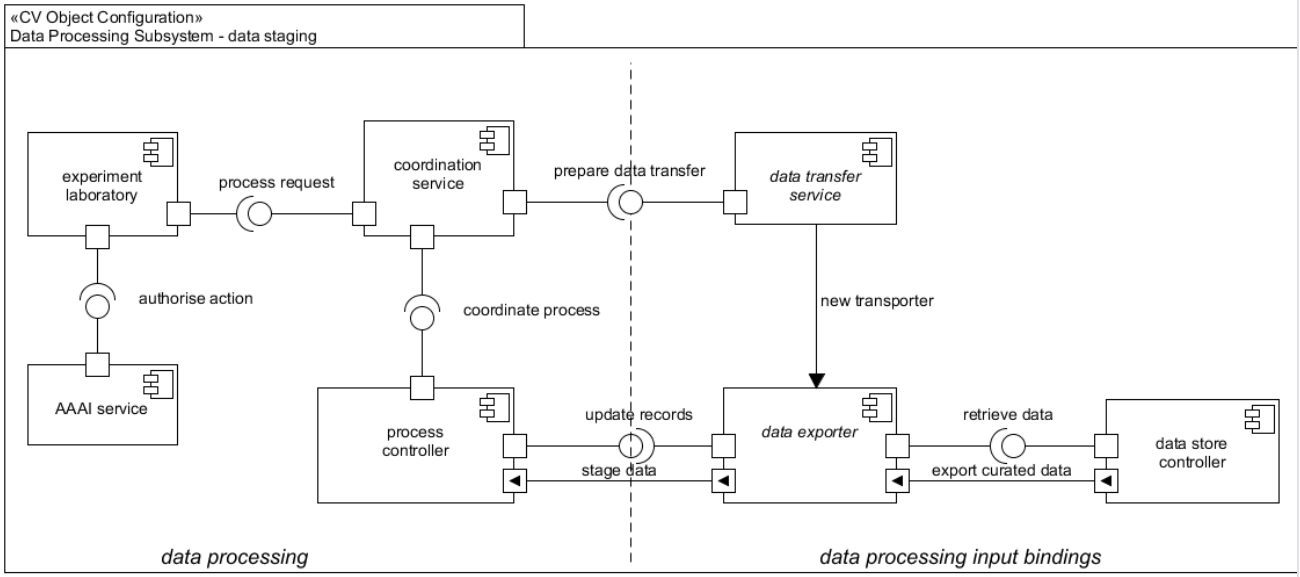
Data Staging

The internal staging of data within an infrastructure for processing requires coordination between data processing components (which handle the actual processing workflow) and data curation components (which hold data within the infrastructure). The diagram bellow displays these two groups of objects which integrate part of the processing subsystem.

Data processing requests generally originate from **experiment laboratory** which validate requests by invoking an **AAAI service** . The **experiment laboratory** will send a process request to a **coordination service** , which interprets the request and starts a processing workflow by invoking the required **process controller**. Data will be retrieved from the data store and passed to the execution platform, the **coordination service** will request that a **data transfer service** to prepare a data transfer.

Data will be retrieved from the data store and passed to the execution platform, the **coordination service** will request that a **data transfer service** to prepare a data transfer. The **data transfer service** will then configure and deploy a **data exporter** which will handle the transfer of data between the storage and execution platforms, i.e. performing data staging. A data-flow is established between all required **data store controller**s and **process controller** via the **data exporter**. After the data-flow is established, processing starts. Processing can include a host of activities such as summarising, mining, charting, mapping, amongst many others. The details are left open to allow the modelling of any processing procedure. The expected output of the processing activities is a derived data product, which in turn will need to be persisted into the RIs data stores.

Data Processing Subsystem - data staging



Notation

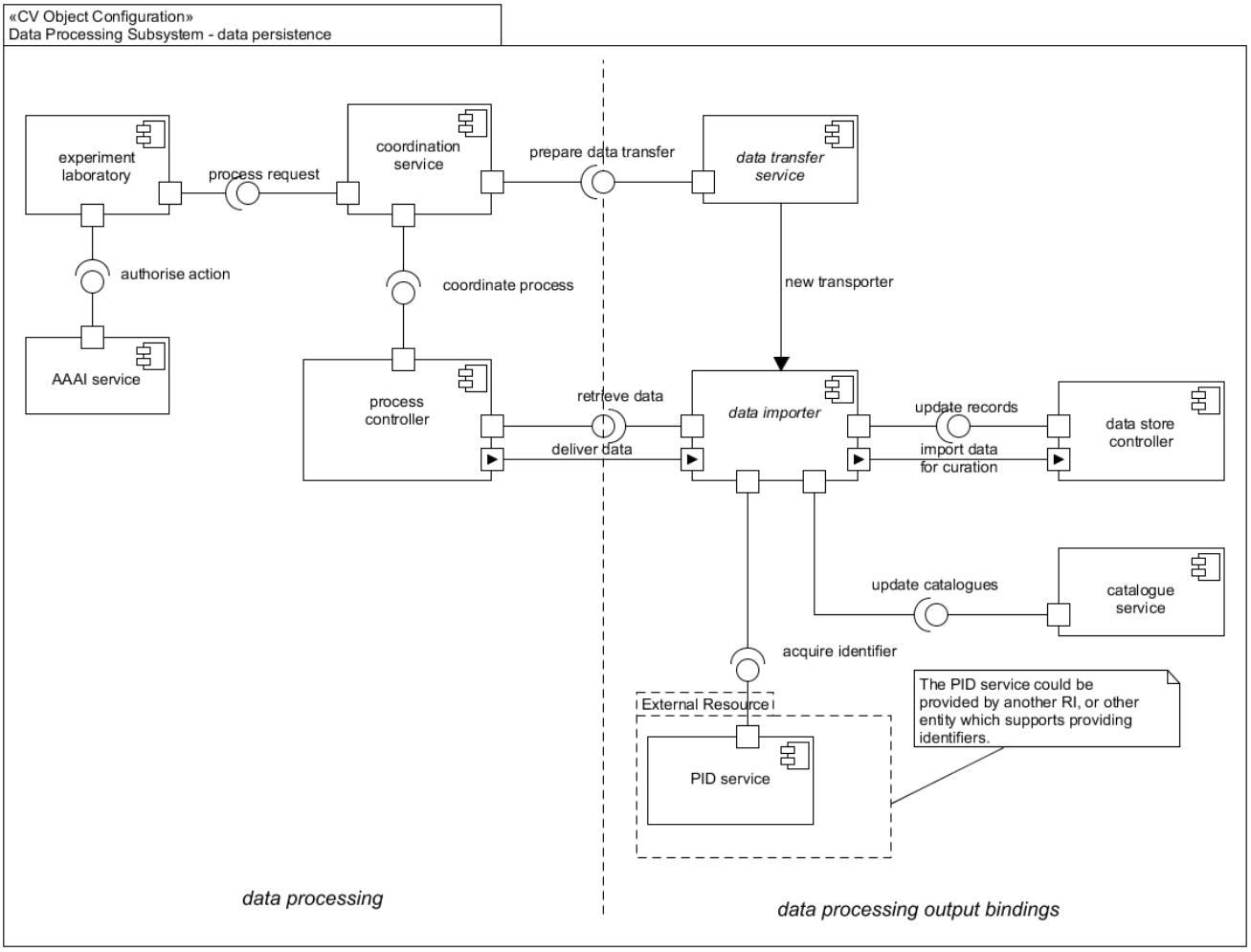
Data Persistence

The persistence of derived data products produced after processing of data within an infrastructure also requires coordination between data processing components (which handle the actual processing workflow) and data curation components (which hold data within the infrastructure). The diagram bellow displays these two groups of objects which integrate part of the processing subsystem.

Data processing requests generally originate from **experiment laboratory** which validate requests by invoking an **AAAI service** . The **experiment laboratory** can present results and ask the user if the results need to be stored, alternatively the user may configure the service to automatically store the resulting data. In either case, after processing, the **experiment laboratory** will send a process request to the **coordination service** , which interprets the request and invokes the **process controller** which will get the result data ready for transfer.

The **data transfer service** will then configure and deploy a **data importer** which will handle the transfer of data between the execution and storage platforms. A data-flow is established between **process controller** and **data store controller** via the **data importer**. After the data-flow is established, the data transfer starts. The persistence if data will trigger various curation activities including data storage, backup, updating of catalogues, requiring identifiers and updating records. These activities can occurs automatically or just as signals sent out to warn human users that an action is expected.

Data Processing Subsystem - data persistence



Notation