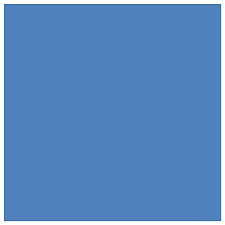
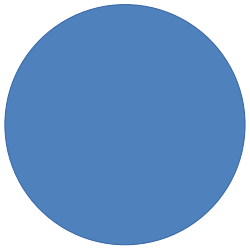
External Entity



Process



Data Store

Data Flow

https://www.lucidchart.com/pages/data-flow-diagram

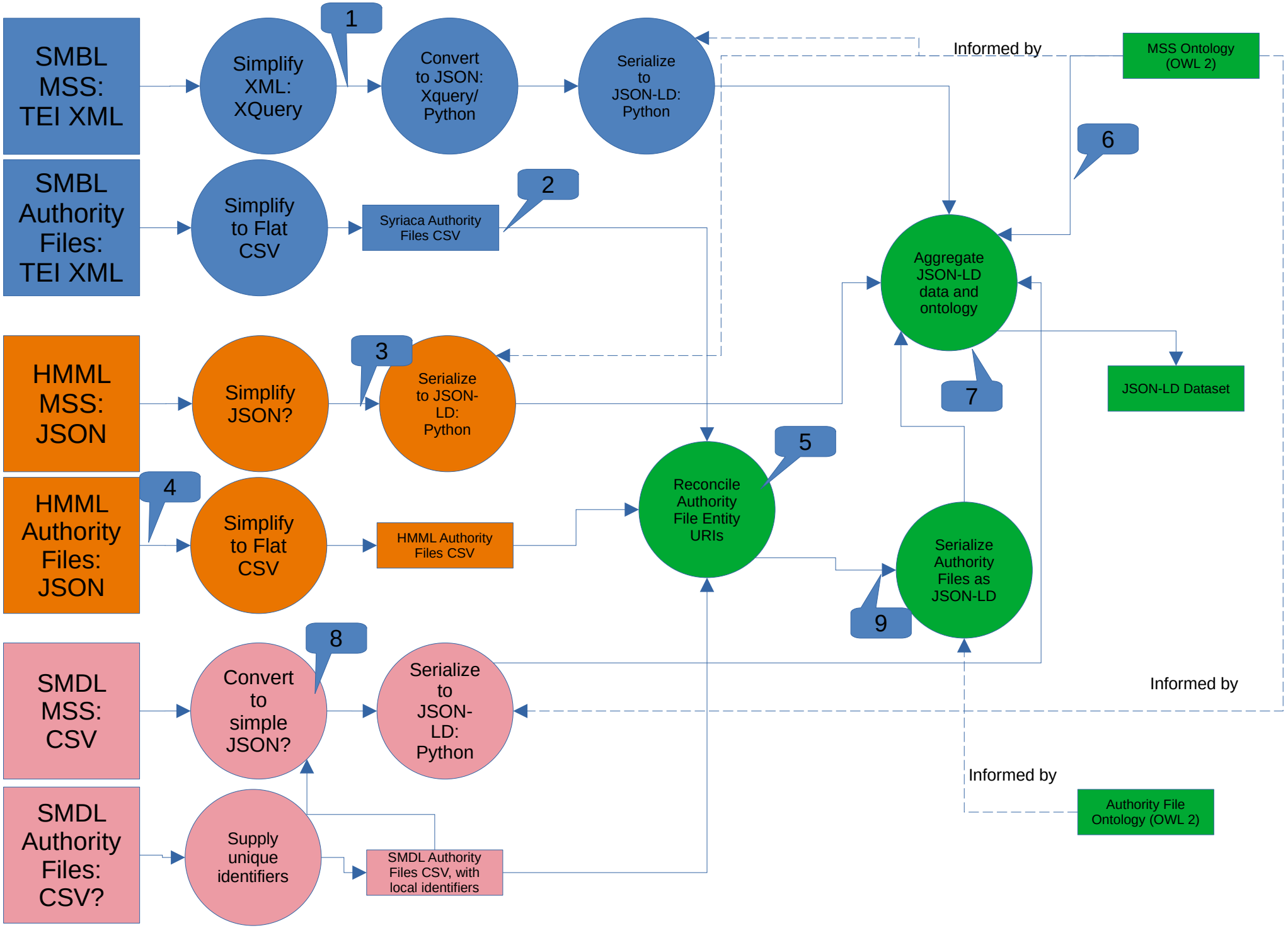
External entity: an outside system that sends or receives data, communicating with the system being diagrammed. They are the sources and destinations of information entering or leaving the system. They might be an outside organization or person, a computer system or a business system. They are also known as terminators, sources and sinks or actors. They are typically drawn on the edges of the diagram.

Process: any process that changes the data, producing an output. It might perform computations, or sort data based on logic, or direct the data flow based on business rules. A short label is used to describe the process, such as “Submit payment.”

Data store: files or repositories that hold information for later use, such as a database table or a membership form. Each data store receives a simple label, such as “Orders.”

Data flow: the route that data takes between the external entities, processes and data stores. It portrays the interface between the other components and is shown with arrows, typically labeled with a short data name, like “Billing details.”

This diagram uses Yourdan and Coad’s notation for data flow components.


SMBL MSS: TEI XML
Simplify XML: XQuery
1
Convert to JSON: Xquery/ Python
Serialize to JSON-LD: Python
Informed by
MSS Ontology (OWL 2)
6
SMBL Authority Files: TEI XML
Simplify to Flat CSV
Syriaca Authority Files CSV
2
Aggregate JSON-LD data and ontology
HMML MSS: JSON
Simplify JSON?
3
Serialize to JSON-LD: Python
7
JSON-LD Dataset
HMML Authority Files: JSON
4
Simplify to Flat CSV
HMML Authority Files CSV
Reconcile Authority File Entity URIs
5
Serialize Authority Files as JSON-LD
9
SMDL MSS: CSV
8
Convert to simple JSON?
Serialize to JSON-LD: Python
Informed by
Informed by
SMDL Authority Files: CSV?
Supply unique identifiers
SMDL Authority Files CSV, with local identifiers
Authority File Ontology (OWL 2)

Notes and Comments:

1: These processes could either be run at once, i.e. the same script both extracts a simplified record for each manuscript *and* converts that flattened record into JSON (i.e., Xquery that exports JSON serialization); or this could be multiple steps, with the intermediate products stored as files. Involved in the simplification process as well might be the removal of non-uniform data that is not easily serialized as JSON-LD (this goes for HMML and SMDL data as well)
2: The relevant parts of an authority file can be expressed as a flat CSV (URI, preferred name, etc.), which will be needed for an entity reconciliation step between the data sets.
3: Further data profiling is required to determine whether initial data preparation/simplification is needed for the HMML dataset or if it can be immediately mapped and serialized to JSON-LD.
4: While the HMML Authority File (HAF) has a robust API for accessing records programmatically, there does not appear to be a publicly-available data dump. It is also the case that the HMML manuscript records reference VIAF URIs rather than HAF ones (for legacy data reasons), so retrieving the authority file data may prove challenging.
5: Many authority file entities (persons, works, places) can be automatically reconciled based on shared URIs. Syriaca URIs are often included in HAF records; likewise both make use of VIAF where possible. Additional reconciliation (manual comparison of names, etc.) may be beyond the scope of this proof of concept project.
6: An important part of the Ontology, besides defining the classes and properties for the JSON-LD will be declaring the Individuals for things like Form (Codex, Quires, Bifolia, etc.); Support Material (Parchment, Paper), and so forth.
7: This may involve minting of project-specific URIs for the manuscript records and associated data? (while retaining sameAs links)
8: The CSV export from SMDL's Airtable includes several fields with multiple 'rolled up' values, so an intermediate JSON representation may simplify the serialization to JSON-LD.
9: It is possible that the serialization of authority files should draw from both a table of reconciled entities and the original format of the authority files. That is, there may be data not easily includable in the CSVs used for entity reconciliation that should be included in the serialization of the JSON-LD. For example, alternate name forms may be too cumbersome for the CSV but easily converted to JSON-LD.