Ingestion Scenarios Draft

- **Create profile for new partner**

Zooey is the coordinator of an aggregation project. He wants to add a new data source (a partner) to his project. He has been in touch with the organization publishing the feed and wants to add the following information from them to the project:

- the type of data source and necessary parameters
  - OAI feed: the feed URL, the metadata prefix, whether sets are enabled and which sets to harvest
  - API: URL and parameters
  - Static data download (for example, an XML file): location and ??
- the official name of the organization as it should appear in records and interfaces
- a primary point of contact for the organization and their contact details
- a secondary point of contact for the organization and their contact details

- **DPLA staff reviews data feed from partner**

Zooey's partner organization now has a feed of data ready to be reviewed before it in aggregated. He has several methods he can use to review this data using tools like pyoaiharvest and Open Refine. After reviewing, Zooey is ready to report back to the partner on changes that are needed to their data and to develop a crosswalk as a guide for writing a mapping.

- **DPLA staff writes mapping**

Zooey translates his crosswalk into a mapping document in the DSL.

- **start a harvest**

Once the crosswalk has been agreed on by the partners and Zooey writes a mapping. Zooey is ready to start a harvest of records from the provider. He logs into the system and first see his dashboard: a page that lists each of his partners and the last activity that was run for each. He selects the new partner (who is listed with no activities, as it was previously added to the system, but has run no activities yet) and goes then to the dashboard page for that provider.

On this page, Zooey can choose to start a harvest. He confirms that the harvest details are correct and chooses to start the harvest. He now sees a status message that tells him the status of the activity, when the harvest started, how many records have been harvested and how many are left to go and how long the harvest has been running.

When the harvest is complete, Zooey sees a confirmation message that the harvest has completed and how many records were downloaded.

- **Harvest fails**

Zooey has started a harvest for a partner and it begins with no error. He sees records start to download and turns his attention elsewhere. He eventually comes back to check on progress

and finds that the harvest has failed. He sees an error message on his screen telling him that the harvest failed at record number 135,764 because of invalid xml. He can now choose whether or not to keep or discard the records already harvested. He chooses to keep the records so that he can test the mapping against them while the partner fixes the error.

- **stop a harvest**

Zooey has started a harvest for a partner, but realizes partway through that he made a mistake in the harvest location. He chooses to stop the harvest process and the process stops. He is asked if he would like to keep or discard the harvested records. Zooey decides to discard the records.

- **Mapping is "added" to the system and deployed**

Zooey has written the mapping for the new partner and has saved it as "CSU_dc.rb." He logs into the system and selects the provider to go to their dashboard. Zooey sees an option to add a new mapping for this provider. He uses the upload tool on the page to add the mapping document.

    - is this based on dashboard or is it backend github stuff?
    - is mapping core code or an asset that is added to the system?

- **mapping is tested**

Zooey decides he wants to test the mapping against a small set of the records. He chooses the option on the screen to test the mapping and chooses a set size of 10% of the records. The system starts mapping the records and a status message tells him that mapping is underway and is updated as records are mapped. When the mapping is complete Zooey sees a confirmation message and a note of how many records were mapped.

    - on a random sample?
    - unsure what that means

- **Mapping is run**

Zooey now wants to run the mapping on the entire downloaded set of records. He chooses the option on the screen to run the mapping against the entire set. The system starts mapping the records and a status message tells him that mapping is underway and is updated as records are mapped. When the mapping is complete Zooey sees a confirmation message and a note of how many records were mapped.

- **Stop mapping**

Zooey starts to run a mapping on a set of downloaded records. However, after he starts it, he realizes that he forgot to upload an updated mapping. He chooses to stop the mapping and the process stops. He is asked if he wants to keep the records that have been mapped so far, or if he wants to discard them. Zooey decides to discard the records.

- **Mapping fails**

Zooey wants to run a mapping on the entire downloaded set of records. He chooses the option on the screen to run the mapping against the entire set. The system starts mapping the records and a status message tells him that mapping is underway and is updated as records are mapped. Zooey turns his attention elsewhere. He comes back to his screen a bit later and finds that the mapping process fails. He sees an error message on the screen that says that the mapping failed on record 134,567. However, despite the individual record fail the mapping continued with other records. Unfortunately, around record 543,210 it ran out of memory and the mapping failed. Zooey sees an error message that tells him this and asks if he wants to keep or discard the already mapped records. Zooey chooses to discard them.

- **Quality control of mapped records**

Once mapping is complete (either a full mapping or a test), Zooey chooses the option to view the records in the quality control view. In the quality control application Zooey is trying to determine if the mapping written for the providers has any errors. He selects his partner from the main page and sees at the top of all subsequent pages the name of the partner, the date of the last index into QA and the number of records. He does the following:

- Looks at the validation reports. These are reports of records that do not contain the required elements (which were set by Zooey as administrator, and are universal across all the partners). Exports a list of records that are missing any of the required elements to send to the partner
- Looks at the metadata property usage report. This is a report that shows every property that appears in any of the mapped records and the number of records the property appears in.
- Uses the facets in the record viewer to look at the languages, place names, creators, and collections to see if anything looks out of place
- Additionally uses any of the values-by-field reports to check for missing or out of place values. These are reports of all of the values for each property (i.e. one report per property). Zooey downloads the ones he is interested in as a CSV file to analyze.
- Scans results in the list view to look for missing or out of place thumbnails
- Finally, randomly browses through side-by-side records checking the mapping of each element in the crosswalk

- **Fixing a mapping and re-running**

Zooey has found a problem in his mapping and needs to revise and re-run it. He downloads a copy of CSU_dc.rb to his desktop to be sure he is working with the correct version of the file. Using notepad, he makes the needed changes to the file. He logs back into the system and navigates to the dashboard for this partner. He uses the option to add a new mapping and uploads the new file. The system asks if he wants to replace the existing mapping and Zooey agrees. His old mapping is saved to the system with the date it was last updated. If Zooey wanted to use this old mapping again, he would be able to choose it when going through the

mapping uploading process. Now he can run a new test or complete re-mapping and go through his QA process again.

- **Setting up an enrichment template**

This is the first time Zooey has used this system and he wants to set up a generic enrichment profile. He navigates to the area of the system where he can do this [?], and he sees an interface that shows him each enrichment module. For each, he can select from a list of properties in the DPLA map. He can also drag and drop to re-order the modules (for example, he wants the DCMI enrichment to run on the formats, normalizing their values, before the punctuation enrichment runs on them to remove punctuation). Once Zooey is satisfied with the profile, he selects “save as template” and he names the template before it saves.

- **Editing an enrichment profile**
    - needs to be thought out a little more. Is it an asset that you upload (profile)? Does it persist? Is it code?
    - add custom enrichment that will be persisted and re-usable

Zooey needs to update the generic enrichment profile for his partner before he can enrich his mapped records. He navigates to the enrichment section of the system for his partner and he selects his generic enrichment template. He then chooses to edit the template and adds the enrichment for truncation of values to the description property. He then saves the custom enrichment profile as “CSU”.

- **Running an enrichment**

Zooey now wants to enrich his mapped records. He navigates to the enrichment section of the system for his partner and chooses to start the enrichment. The system starts enriching the records and a status message tells him that enrichment is underway and is updated as records are enriched. When the enrichment is complete Zooey sees a confirmation message and a note of how many records were enriched.

- **enrichment fails**

Zooey now wants to enrich his mapped records. He navigates to the enrichment section of the system for his partner and chooses to start the enrichment. The system starts enriching the records and a status message tells him that enrichment is underway and is updated as records are enriched. Unfortunately, the system had a timeout error while the process was running. Zooey returns to his screen and sees an error message that says that this error occurred and asks him if he wants to keep or discard the already enriched record. Since most of the records were enriched, Zooey decides to go ahead and keep them for now. He will handle the remaining records in a subsequent ingestion.

- **enrichment is stopped**

Zooey now wants to enrich his mapped records. He navigates to the enrichment section of the system for his partner and chooses to start the enrichment. The system starts enriching the

records and a status message tells him that enrichment is underway and is updated as records are enriched. Zooey realizes after the enrichment starts that he forgot to add another custom enrichment to the profile. He chooses to stop the enrichment. He is asked if he wants to keep the records that have been enriched so far, or if he wants to discard them. Zooey decides to discard the records.

- **Quality control & reports of enrichment**

Zooey is now ready to do QA of the enriched records. His methods are very similar to the previous QA, but he additionally looks at validation reports of the type property and pays particular attention to the spatial and subject facets because those values have now been enriched. When Zooey is satisfied with what he sees, he sends notification to the partner that they may do some QA analysis themselves using the same QA app (they will create a username and password for themselves). This step ends when either both partners are happy with what they see, or an error is found that needs to be fixed.

- **running an additional enrichment**

Zooey has realized that he failed to run the Genre enrichment. He knows that this enrichment can be simply be run on the already enriched records because it affects a property that was unchanged during the previous enrichment. He navigates to the enrichment part of the system and selects “run additional enrichment”. From there he chooses to the run the genre enrichment on the hasType property on the enriched records. The system asks him if he would like to add this to the enrichment profile for future enrichments. Zooey indicates yes. The enrichment runs on the records. and a status message tells him that enrichment is underway and is updated as records are enriched. When the enrichment is complete Zooey sees a confirmation message and a note of how many records were enriched. Once the process is complete Zooey opens the QA app to check the new records.

- **fix and re-run an enrichment**

Zooey has realized that he ran the type enrichment on the wrong property. This has resulted in changing the values from the mapped records incorrectly. In order to fix this Zooey will have to re-map the records, fix the enrichment profile, and then re-enrich them. Fortunately, the enrichment section of the system contains an option for this. Zooey selects “re-map and fix enrichment” and selects the ”CSU” enrichment profile he previously saved. The enrichment profile opens to the same screen as before and Zooey corrects the type enrichment. He hits “save” and is asked if he wants to replace the saved version or create a new file. He chooses to replace the saved version of the profile. Next he is asked to select the harvest set of records to be re-mapped and enriched. Next, he is asked if he wants to use the current mapping “CSU_dc.rb” or create a new one. He chooses to use the existing mapping. Finally, Zooey is asked if he wishes to run the mapping and enrichment now. He selects yes and the mapping starts. The system starts mapping the records and a status message tells him that mapping is underway and is updated as records are mapped. When the mapping is complete Zooey sees a confirmation message and a note of how many records were mapped. Then, the enrichment of

records starts. The enrichment runs on the records. and a status message tells him that enrichment is underway and is updated as records are enriched. When the enrichment is complete Zooey sees a confirmation message and a note of how many records were enriched. When all the processes are complete, Zooey starts the QA process again.

- **check on staging**

When both the partner and Zooey are happy with the records in the QA interface, Zooey is ready to index the records to the staging server. In the system, on the partner dashboard he chooses to index the enriched records to the staging server. Zooey is notified that the indexing has started and is updated as records are indexed. A status message tells him when the indexing is complete. Zooey navigates to the version of the DPLA portal on staging and reviews the records. In this view of the record he can check specific quality issue that would only surface in the portal such as how the records appeared in the timeline or the map browse. When he is satisfied with how the records look, he notifies the partner to view the records. This step ends when both parties are happy with the records.

- **publish**

When both the partner and Zooey are happy with the records in staging, Zooey is ready to index the records to the production server. In the system, on the partner dashboard he chooses to index the enriched records to the production server. Zooey is notified that the indexing has started and is updated as records are indexed. A status message tells him when the indexing is complete.

- **re-ingest**

Zooey is ready to re-ingest records from a partner. He navigates to the partners dashboard and selects “Do a reharvest.” The system begins by harvesting new records. A status message tells him that harvesting has begun and gives him updates as records are downloaded. When the harvest is complete, Zooey sees a status message that tells him how many records were added, how many were changed, and how many were deleted, and the new total from that harvest. Since everything looks good to him, Zooey selects the option to map and enrich the records. The system starts mapping the records and a status message tells him that mapping is underway and is updated as records are mapped. When the mapping is complete Zooey sees a confirmation message and a note of how many records were mapped. Then, the enrichment of records starts. The enrichment runs on the records. and a status message tells him that enrichment is underway and is updated as records are enriched. When the enrichment is complete Zooey sees a confirmation message and a note of how many records were enriched. When all the processes are complete, Zooey starts the QA process. He does a quick check of the records in the QA interface, checking some of the reports and facets to make sure that nothing looks incorrect. When he is satisfied everything is good, he navigates back to the dashboard and chooses to index the enriched records to the production server. Zooey is notified that the indexing has started and is updated as records are indexed. A status message tells him when the indexing is complete.