Staging Area Design

Definitions

Live Objects: Dataset and Collections in a Space. Anybody can change/add/modify.

Curation Object: Object being modified for publication by the curators.

Publication Object: Curation Object + DOI (Digital Object Identifier). The returned object from the repository.

Process

1. Curation Area - Curator Picks some of the Live Objects and start the process of submitting that dataset or collection.

    Design Specification:

    - The user should be able to identify the datasets and collections that they want to publish and create a curation object.
    - There is a curation Area specific for each space.
        - Anyone with correct permissions in the space can see the curation area and objects in it.
        - Private staging area will exist as well, to allow people to push out items that they haven't shared
        - While in the staging area, each curation object will have its own flow, independent of any other curation object

    Design Question: What is a curation object, with respect to Clowder?

    Design Answer: It is an object that holds copies of the live objects. The curation object contains what was selected from the live objects and is put into the staging area as an object that the user will be working on. There is an entry in the staging area with all the information that was available when the user added it to the staging area.

    Design Question: What happens if the original dataset/collection is updated? Do we keep track in the curation dataset to the live object?

    Design Answer: The curation Dataset/Collection has a link to the original dataset/collection. They will have the same Id, but stored in different places. This link should be bi-directional as the original Live Objects will have a reference to the Curation Object that is in process, or the Publication Object, after the process is finished.

    Design Question: There will be a staging area per space. Is there also a staging area that is private?

    Design Answer: Pending

2. Matchmaker - The user 'asks' the program what repositories will accept the curation object.

    Design Specification

    - Identify what repositories are compatible with what the users wants to do. (I am interested in these things for my publication. What repositories are compatible with this requirements?)
    - User has choices for some available, and selection boxes/options regarding requirements, and matchmaker comes back with options.
        - Clowder should save the preferences for the repositories the user wants to publish to. Clowder should track/remember that for future publication flows (probably stored in profile)
        - Need to study and determine how matchmaker takes those rules in.
            - Input to matchmaker: User Preferences + attributes of the data.
    - As the choices come back, is there information/feedback to guide refinement?
        - As an example, if one place needs an abstract, staging area will provide GUI to add that in at this step in the flow.
        - No need to build this type of validation/confirmation infrastructure yet. Sometimes that feedback happens out-of-band.

    Design Question: If, during the curation flow, the live object is updated and I want to update my curation object, can this be done or should it be done? Or is it best to just force them to start a new flow.

    Design Answer: Initial answer was to make them start over. After some discussion and later in the conversation, that seems to have changed, so To Be Determined? It should be noted that these curation objects can be long-lived and that could mean weeks or possibly months.

3. C3PO

    Design Specification.

    - Find a way for helping the user refine what the repository needs for allowing a submission. Example: If the repository requires an abstract. Provide a way for the user to add an abstract to the dataset. Or other kind of metadata/requirements for the dataset/collections in order for the curation object to be accepted for publication.

TO BUILD:

Discussion Points:

- Plugins to handle these. 1) Staging Area Plugin 2) Ideals Plugin 3) C3PR plugin?
    - Modularity will allow for incremental/iterative building as well as making it easier for future extensibility in terms of recommenders (matchmaker) and publishing work (Ideals, C3PR, etc...)

- During Sprint 1), should also be learning how to submit to matchmaker and C3PR so that can help make sprint 2) easier, where the interface will be built for those items to integrate into the flow.
- Eventually able to edit, up until a certain step in the flow. To start, no edit?
- Question - metadata exists on the curation object? Yes, probably. How in depth? Basic items, curator, date, title, desc

Sprint 1 - Curation Projects.

- Create a staging area per space
- Create a curation object: creator, date, title, description, curator, dataset, collections, list of copies, blob, metadata, object
- Store users preferences for publication

Optional:

- Edit Curation Object

Sprint 2 - Matchmaker Calls

- Store user preferences for publication. (Preferences for calling the matchmaker)
- Add rules to the matchmaker that are attributes of the data.
- Wizard Flow
- Call Matchmaker with user preferences + attributes of the data

Sprint 3 - C3PR Calls.

- Refine Metadata - Show suggestions to the user. About missing information that the repository needs/wants. Optional
- Create multiple plugins for submitting the code a C3PR plugin, Ideals plugin fedora plugin.

Sprint tasks:

1. Staging area per space [Indira]
    a. standalone plugin
2. Create curation object [Yan]
    a. Select dataset and collections from space [Luigi]
3. Submit for publication (separate plugins?) [Rob]

1. Call matchmaker and pick repository (separate plugins?)
2. Refine metadata
3. Store user preferences for publication in profile
4. Store published object (everyone in space can see them)

1. Edit curation object
2. List curation objects and published objects that a dataset/collection are part of

Steps

1. Create curation object
2. Matchmaker query and selection
3. Editing of metadata and submission to repository

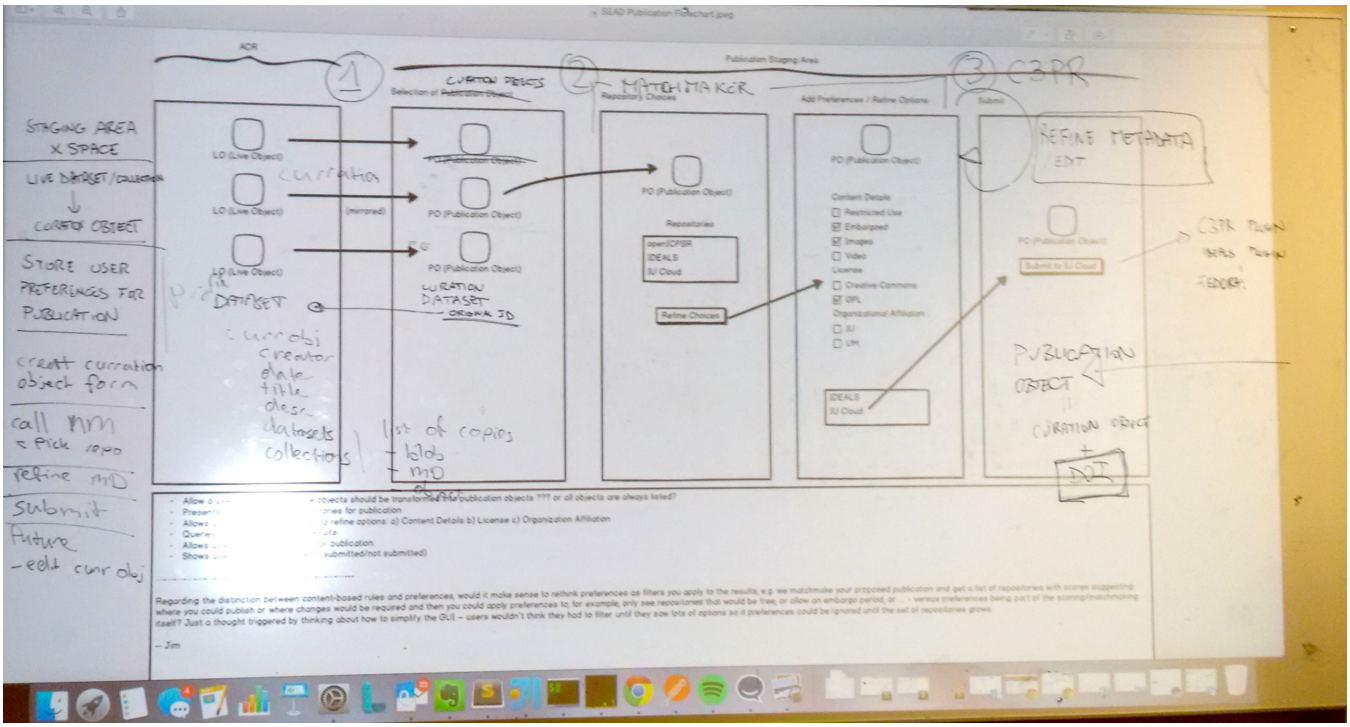
Questions

1. Who can see the curation objects?

Background

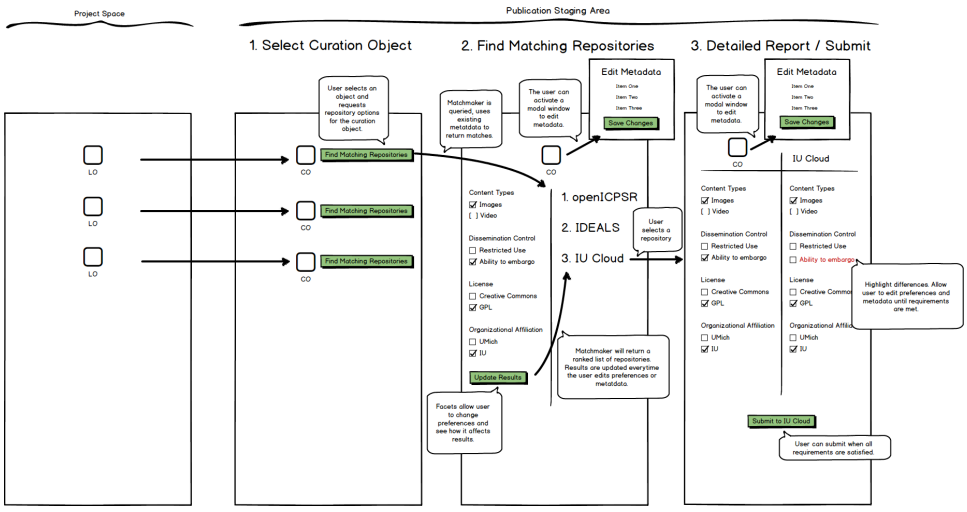
1. curation object -> publication object
2. repositories preferences
    a. a repository says what options it provides
    b. user might have preferences set in their profiles in the spaces
        i. generally speaking I want things free
        ii. but in one instance I might be willing to pay
3. attributes of the content vs attributes of the repositories
    a. "I would like"
    b. "I have images"
4. "if my dataset doesn't have license, assign creative common"
5. preference / requirements

Board Picture



Reference Pages

- Wiki page for design: WP-4.4 Staging Area UI design - WP 5.2: SEAD 2.0 API/Shim Design Sprint
- Michael's Flow diagram:



Sample Rules from Inna

1. "Affiliation match – string match

2. Max collection hierarchy - # subfolders / 1 means flat structure

3. Max collection size – number in GB / no limit

4. Max file size – number in GB / no limit

5. Preferred formats – list / any (if list, also need to check whether conversion is an option)

6. Metadata required – none / structured (fields required) / un-structured (data description or readme) / both required

7. Domain match – string match

8. Packaging - preferred type (zip, tar, gz, bagit) / packaging not preferred

9. Versions accepted – new submissions, updates, special versions, all

10. Confidential info in data – acknowledgement required / not required

11. Copyrighted material – acknowledgement required / not required

12. Data License – required / not required (type - repository specific use agreement / any type)

13. Depositor Agreement – required / not required (need to prepared or accepted at the time of matchmaking?)

14. Access – open, restricted, embargo, enclave

It seems to me that at least some of this information would be generated in the staging area via clicking on some checkboxes. Some of it will come from project spaces metadata fields, but I'm not sure we have appropriate predicates. I'm still unclear about how to find them. I saw that all the rules that already exist have "http://sead-data.net/terms/" as part of their implementation. Does it mean we'll rely on our own vocabulary for all the rules?"

Curation Statuses

- Under review for publication
- Submitted (RO submitted to publication, but is still in SEAD)
- Transferred (RO transferred to repository)
- Under review (repository is reviewing the package for completeness, etc.)
- In curation (repository is curating the RO)
- Approved / rejected for publication (repository made the decision about publishing RO)
- Published / returned (RO published or returned to SEAD if it was rejected)

Other curations: does it open in an app? is there a code book? are variables named properly? Rearrange folders

---

09/02/2015 Notes:

Discussion:

1) Only the owner of a dataset or a member of a space with the EditStagingArea Permission can see the publish button for a dataset.

2) Discussion about when a dataset is selected for publication. The curation object associated with the dataset is displayed on the staging area for all the spaces the dataset is in.

3) Discussed a new possible flow for the application. Where the flow starts with the submit page with a pre/selected repository and the matchmaker results displayed. The preselected repository can be either the user's preferred repository or the last repository used. The page could have a message saying "Please select a repository" if none of the above cases is possible. There will be a button next to the repository name to change the repository which will lead to the current Matchmaker page. The Edit Metadata page will be accessible via the link in left navigation. A validation button should also be available in the main page to verify that the dataset is ready for publication.

Reference API

https://sead-test.ncsa.illinois.edu/sead-cp/

https://github.com/qqmyers/sead2