Datasets Usage Metadata Extraction from OpenAlex

This document outlines the process of extracting datasets usage metadata from OpenAlex and storing it in a database, enabling decision-makers to utilize this data effectively in the future.

Assumptions

- **OpenAlex API**: The metadata will be extracted using the OpenAlex API.
- **Data Schema**: A simple, initial data schema will be developed to store the extracted metadata, containing only the necessary fields for creating initial dashboards.
- **Raw Data Storage**: Each entity in the data schema will include a property to store the raw data extracted from the OpenAlex API, allowing for future reprocessing if needed.
- **Full-Text Search**: OpenAlex enables full-text searches on publications without requiring prior full-text downloads. This feature will be used for dataset alias string searches.
- **Open Access Publications**: Publications with full-text downloads available, particularly open access ones, will be considered for future machine learning model execution in the next phase of the project.

Data Schema Entities and Properties

The following data schema outlines the properties necessary for creating an initial dashboard. The data will be extracted from the API's responses to standardize the storage of the information needed for dashboard development. All entities will have a `raw` property containing the API response from which the data was extracted.

Collection	Fields	Description
publication	`_id`, `publication_id`, `publication_external_id`, `title`, `year`, `doi`, `citation_count`, `journal_id`, `open_access_url`	Stores publication details and references other entities by their IDs.
journal	`_id`, `journal_id`, `journal_external_id`, `name`, `issn`	Stores information about journals.
publication_author	`_id`, `publication_id`, `author_id`	Maps publications to their respective authors.
author	`_id`, `author_id`, `author_external_id`, `name`, `orcid`	Stores author information.
author_institution	`_id`, `author_id`, `institution_id`	Maps authors to their respective institutions.
institution	`_id`, `institution_id`, `institution_external_id`, `name`, `state`, `country`, `ror`	Stores institution details.
publication_topic	`_id`, `publication_id`, `topic_id`	Maps publications to their respective topics.
topic	`_id`, `topic_id`, `topic_external_id`, `name`, `type`	Stores topic information.
dataset	`_id`, `dataset_id`, `name`	Stores dataset information.
dataset_alias	`_id`, `alias_id`, `dataset_id`, `name`	Stores dataset aliases.
publication_dataset_alias	`_id`, `publication_id`, `dataset_alias_id`	Maps publications to their respective dataset aliases.

OpenAlex Data Extraction

- **Dataset Example**: The NAIRR Pilot Dataset USDA Census of Agriculture will be used.
- **Alias Search**: The following aliases will be used for full-text string searches: "NASS Census of Agriculture", "Agricultural Census", "USDA Census", "AG Census". These aliases have been used to create the seed corpus for the Democratizing Data Research using Scopus.

Collection	Schema Field	Prerequisite	URL	API Field
publication	publication_external_id	dataset_alias	OpenAlex API	`id`
publication	title	dataset_alias	OpenAlex API	`title`
publication	year	dataset_alias	OpenAlex API	`publication_year`
publication	doi	dataset_alias	OpenAlex API	`doi`

Collection	Schema Field	Prerequisite	URL	API Field
publication	citation_count	dataset_alias	OpenAlex API	`cited_by_count`
publication	open_access_url	dataset_alias	OpenAlex API	`open_access.oa_url`
journal	journal_external_id	dataset_alias	OpenAlex API	`[locations[0]|best_oa_location].source.id`
journal	issn	dataset_alias	OpenAlex API	`[locations[0]|best_oa_location].source.issn_l`
journal	name	dataset_alias	OpenAlex API	`[locations[0]|best_oa_location].source.display_name`
author	name	dataset_alias	OpenAlex API	`authorships[*].author.display_name`
author	orcid	dataset_alias	OpenAlex API	`authorships[*].author.orcid`
author	author_external_id	dataset_alias	OpenAlex API	`authorships[*].author.id`
institution	name	dataset_alias	OpenAlex API	`authorships[*].institutions.display_name`
institution	ror	dataset_alias	OpenAlex API	`authorships[*].institutions.ror`
institution	country	dataset_alias	OpenAlex API	`authorships[*].institutions.country_code`
institution	institution_external_id	dataset_alias	OpenAlex API	`authorships[*].institutions.id`
institution	state	dataset_alias|institution|ror	ROR API	`addresses[0].state`
topic	topic_external_id	dataset_alias	OpenAlex API	`topics[*].id`
topic	name	dataset_alias	OpenAlex API	`topics[*].display_name`