Week 1,2,3

Intended work

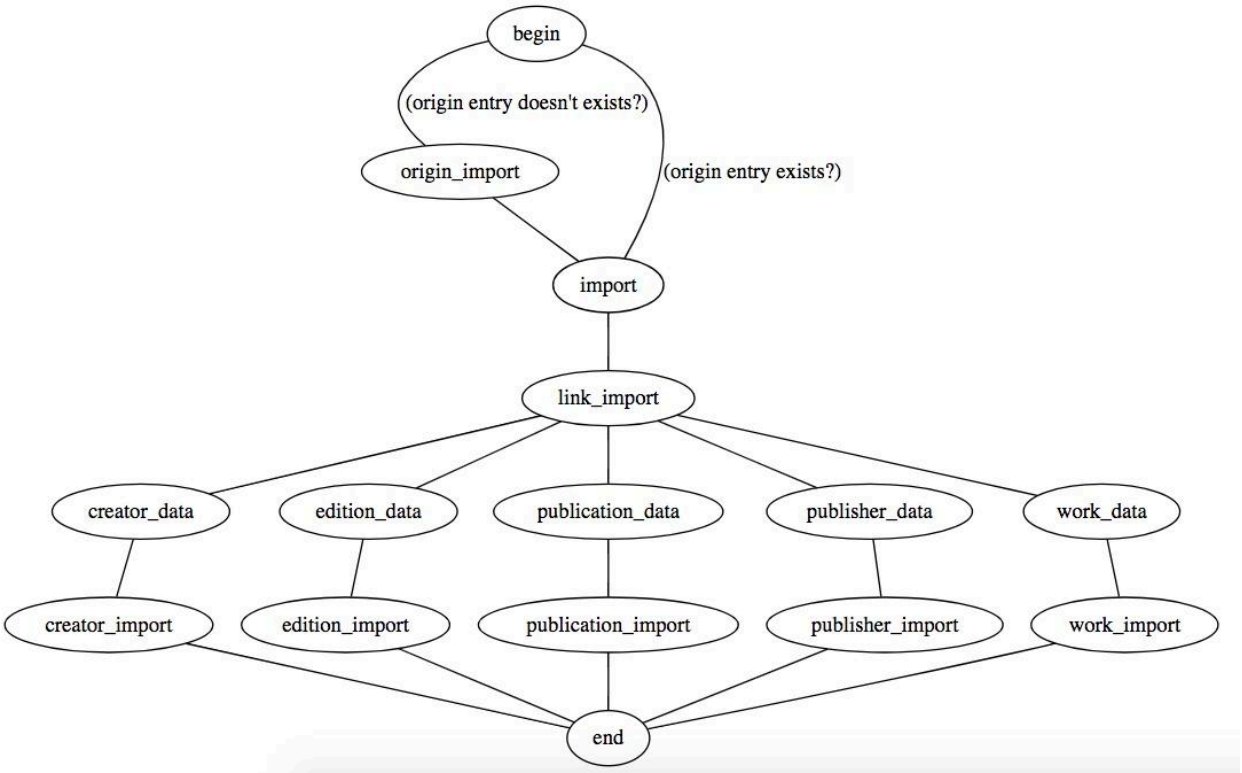
1. Write SQL queries reflecting schema changes in bookbrainz-sql.
    a. Create relevant tables which primarily include:
        i. `bookbrainz.creator_import;`
        ii. `bookbrainz.edition_import;`
        iii. `bookbrainz.publication_import;`
        iv. `bookbrainz.publisher_import;`
        v. `bookbrainz.work_import;`
        vi. `bookbrainz.discard_votes;`
        vii. `bookbrainz.link_import;`
        viii. `bookbrainz.origin_import;`
        ix. `bookbrainz.import;`
2. Propose a generic data object to be developed from the dumps to be used later by import script.This generic data object is filled in with other relevant information specific to the import object (and not the source).
    a. Decide actions of consumers and producers and the kind of object they will be processing.
3. Write relevant functions to access the newly created import object in bookbrainz-data.
    a. Insert function - knex queries to provide a single function abstraction to add an import which will handle all relevant sub-functions.
    b. This would primarily include:
        i. Updating the origin_import, import table and link_import tables.
        ii. Updating the entityType_data table (ensuring all pre-planned activities like keeping some compulsory portions null via insertion code in some fields like publications and editions). This same function could also be used in the present model of entity addition into entityType_data table (with non-nullable fields staying that way).
        iii. Setting up delete import management functions.
        iv. Setting up discard vote management system. Ideally a function exists which would check the number of votes cast to discard till date and if it is below a certain threshold - a new row is added else the command to delete import and delete all votes in the discard vote system.
        v. Set up a testing environment for the same (?How to do it without seed data?)
4. Analyze the dumps and draw a basic layout of the OL data dumps format.

5. Prune the dumps to a manageable size, while maintaining diversity.
    a. Split the dumps and analyze them one by one to create an ideal testing data set.
    b. Set up an testing environment for imports in bb-data using the above relevant data.
    c. Add handcrafted varying data to provide for various loopholes the data could have.
    d. Write some preliminary tests.

NOTE: I intend to finish them at most within 1.5 weeks and at best in a week.

Follow up

1. Relevant changes have been made in the following PR: https://github.com/bookbrainz/bookbrainz-sql/pull/8
2. Accessor function flow:



- First we check if the source of the import exists in our records, if yes the we proceed to create an import object, fill up the link_import table and then try to insert new records.

- Another way could be ensure that the import objects are only from fixed predetermined sources only, and so first we insert into origin_import table and then carry on.
- Another method could be to get rid of the origin_import table and instead of storing the ids referencing origin_import we directly store the origin name. This however would beat the above point as there would be no check on how to ensure only a fixed origin sources are added to the tables as imports.

3. Object structure for entity type addition on the bb-site are as given below (the access functions in bb-data take in exactly similar objects with slight modification like extra fields etc.):

Creator

```
// Creator specific data


{
   "aliases": [
       {
           "default": true,
           "languageId": 100,
           "name": "test27",
           "primary": true,
           "sortName": "test27"
       },
       {
           "default": false,
           "languageId": 100,
           "name": "test27(b)",
           "primary": false,
           "sortName": "test27(b)"
       }
   ],
   "beginAreaId": 220,
   "beginDate": "1995-12-02",
   "disambiguation": "Dansish test27",
```

	"endAreaId": 225,
	"endDate": "1999-12-02",
	"ended": true,
	"genderId": 2,
	"identifiers": [
	{
	"typeId": 2,
	"value": "189002e7-3285-4e2e-92a3-7f6c30d407a2"
	}
	],
	"note": "test27",
	"typeId": 1
	}

Publisher	// Publisher related data
	{
	"aliases": [
	{
	"default": true,
	"languageId": 98,
	"name": "test28",
	"primary": true,
	"sortName": "test28"
	},
	{
	"default": false,
	"languageId": 18,
	"name": "test28(b)",
	"primary": false,
	"sortName": "test28(b)"

	`}`
	`],`
	`"areaId": 99,`
	`"beginDate": "1997-12-03",`
	`"disambiguation": "test28",`
	`"endDate": "1998-12-04",`
	`"ended": true,`
	`"identifiers": [`
	`{`
	`"typeId": 20,`
	`"value": "Q42"`
	`}`
	`],`
	`"note": "test28",`
	`"typeId": 1`
	`}`

Work	`// Work related data`
	`{`
	`"aliases": [`
	`{`
	`"default": true,`
	`"languageId": 98,`
	`"name": "test29",`
	`"primary": true,`
	`"sortName": "test29"`
	`},`
	`{`
	`"default": false,`

	"languageId": 76, "name": "test29(b)", "primary": false, "sortName": "test29(b)" }], "disambiguation": "test29(b)", "identifiers": [], "languages": [52], "note": "test29", "typeId": 2 }

Publication	// Publication related data
	{
	"aliases": [
	{
	"default": true,
	"languageId": 100,
	"name": "test30",
	"primary": true,
	"sortName": "test30"
	},
	{
	"default": false,
	"languageId": 113,
	"name": "test30(b)",

	"primary": true, "sortName": "test30(b)" }], "disambiguation": "test30 disambiguation", "identifiers": [], "note": "test30", "typeId": 3 }

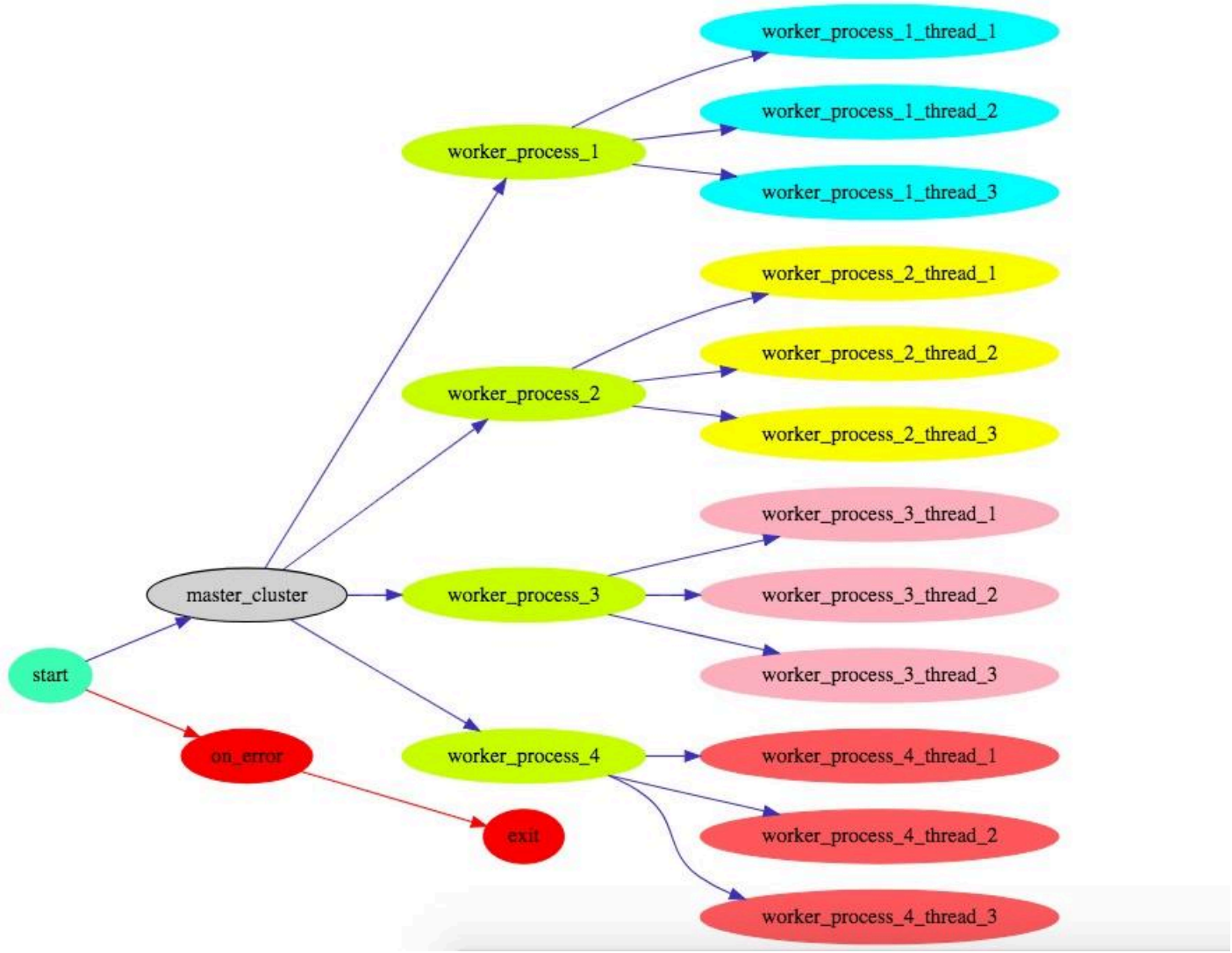
Edition	// Edition related data
	{
	"aliases": [
	{
	"default": true,
	"languageId": 18,
	"name": "test31",
	"primary": true,
	"sortName": "test31"
	},
	{
	"default": false,
	"languageId": 18,
	"name": "test31(b)",
	"primary": true,
	"sortName": "test31(b)"
	}
	],
	"depth": 54,

	"disambiguation": "test31 disambiguation",
	"formatId": 5,
	"height": 23,
	"identifiers": [],
	"languages": [
	18
	],
	"note": "test31",
	"pages": 34,
	"publicationBbid": "c1a7db3a-c1d3-4e4a-bcb3-4bcb0d855843",
	"publishers": [
	"364adcae-6d8b-426c-91d7-ec3fe02c9142"
	],
	"releaseEvents": [
	{
	"date": "1997-12-02"
	}
	],
	"statusId": 2,
	"weight": 12,
	"width": 12
	}

editor	{ "metabrainzUserId": 1977131, "id": 1171, "name": "bukwurm", "reputation": 0, "bio": "", "birthDate": null,

```
    "createdAt": "2018-01-16T14:04:58.919Z",
    "activeAt": "2018-01-16T14:04:58.919Z",
    "typeId": 1,
    "genderId": 1,
    "areaId": null,
    "revisionsApplied": 50,
    "revisionsReverted": 0,
    "totalRevisions": 50,
    "cachedMetabrainzName": "shivamt",
    "titleUnlockId": null
}
```


4. Async Cluster structure:
    - Cluster head (master process) *require('cluster')*
    - Worker processes (= number of CPUs) *require('async')*
        - Threads picking up one dump sub-file.



5. Exploration of dumps:

```
// Authors
{
   "count":  6924508,
   "keys": [
       "name",
       "personal_name",
       "last_modified",
       "key",
       "type",
```

```
        "revision",
        "birth_date",
        "created",
        "death_date",
        "latest_revision",
        "remote_ids",
        "photos",
        "title",
        "bio",
        "links",
        "wikipedia",
        "alternate_names",
        "comment",
        "fuller_name",
        "date",
        "website",
        "location",
        "entity_type",
        "photograph",
        "numeration",
        "create",
        "role",
        "source_records",
        "ocaid",
        "works",
        "body",
        "number_of_pages",
        "lc_classifications",
        "genres",
        "languages",
```

```
        "subjects",
        "publish_country",
        "title_prefix",
        "oclc_numbers",
        "by_statement",
        "publishers",
        "authors",
        "publish_places",
        "pagination",
        "lccn",
        "publish_date",
        "website_name",
        "tags",
        "dewey_decimal_class",
        "notes",
        "subject_place",
        "covers",
        "series",
        "edition_name",
        "id_librarything",
        "id_wikidata",
        "id_viaf",
        "other_titles",
        "subtitle",
        "subject_time",
        "contributions",
        "     _date",
        "marc"
    ]
}
```

```
// Works
{
    "count": 1080593,
    "keys": [
        "title",
        "created",
        "last_modified",
        "latest_revision",
        "key",
        "authors",
        "type",
        "revision",
        "covers",
        "subjects",
        "subtitle",
        "subject_places",
        "subject_people",
        "description",
        "subject_times",
        "cover_edition",
        "links",
        "works",
        "lc_classifications",
        "first_publish_date",
        "dewey_number",
        "first_sentence",
        "excerpts",
        "number_of_editions",
```

```
        "remote_ids"
    ]
}
```

```
// Editions
{
    "count": 25584282,
    "keys": [
        "identifiers",
        "subtitle",
        "subject_place",
        "lc_classifications",
        "latest_revision",
        "contributions",
        "edition_name",
        "title",
        "languages",
        "subjects",
        "publish_country",
        "by_statement",
        "type",
        "revision",
        "other_titles",
        "publishers",
        "last_modified",
        "key",
        "authors",
        "publish_places",
        "pagination",
        "dewey_decimal_class",
```

```
        "notes",
        "number_of_pages",
        "lccn",
        "isbn_10",
        "publish_date",
        "works",
        "physical_format",
        "weight",
        "isbn_13",
        "physical_dimensions",
        "covers",
        "created",
        "oclc_numbers",
        "classifications",
        "series",
        "source_records",
        "subject_time",
        "table_of_contents",
        "ocaid",
        "title_prefix",
        "ia_box_id",
        "description",
        "first_sentence",
        "genres",
        "isbn_invalid",
        "work_title",
        "translated_from",
        "translation_of",
        "coverimage",
        "location",
```

```
        "ia_loaded_id",
        "scan_records",
        "scan_on_demand",
        "work_titles",
        "uri_descriptions",
        "uris",
            "url",
        "oclc_number",
        "isbn_odd_length",
        "links",
        "contributors",
        "copyright_date",
        "full_title",
        "download_url",
        "purchase_url",
        "language",
        "collections",
        "original_isbn",
        "subject_times",
        "subject_places",
        "subject_people",
        "openlibrary",
        "author_names",
        "create",
        "remote_ids",
        "volumes",
        "isbn",
        "language_code",
        "name",
        "birth_date",
```

```
        "numer_of_pages",
        "bookweight",
        "library_of_congress_name",
        "body",
        "vorks",
        "m",
        "lc_classification",
        "macro",
        "price",
        "error",
        "code",
        "edition",
        "ia_id",
        "volume_number",
        "dimensions",
        "coverid",
        "by_statements",
        "dewry_decimal_class",
        "stats",
        "news"
    ]
}
```

6.