McGill

DDMAL DISTRIBUTED DIGITAL MUSIC ARCHIVES & LIBRARIES LAB

Data Processing and Reconciliation

LinkedMusic Project Meeting II

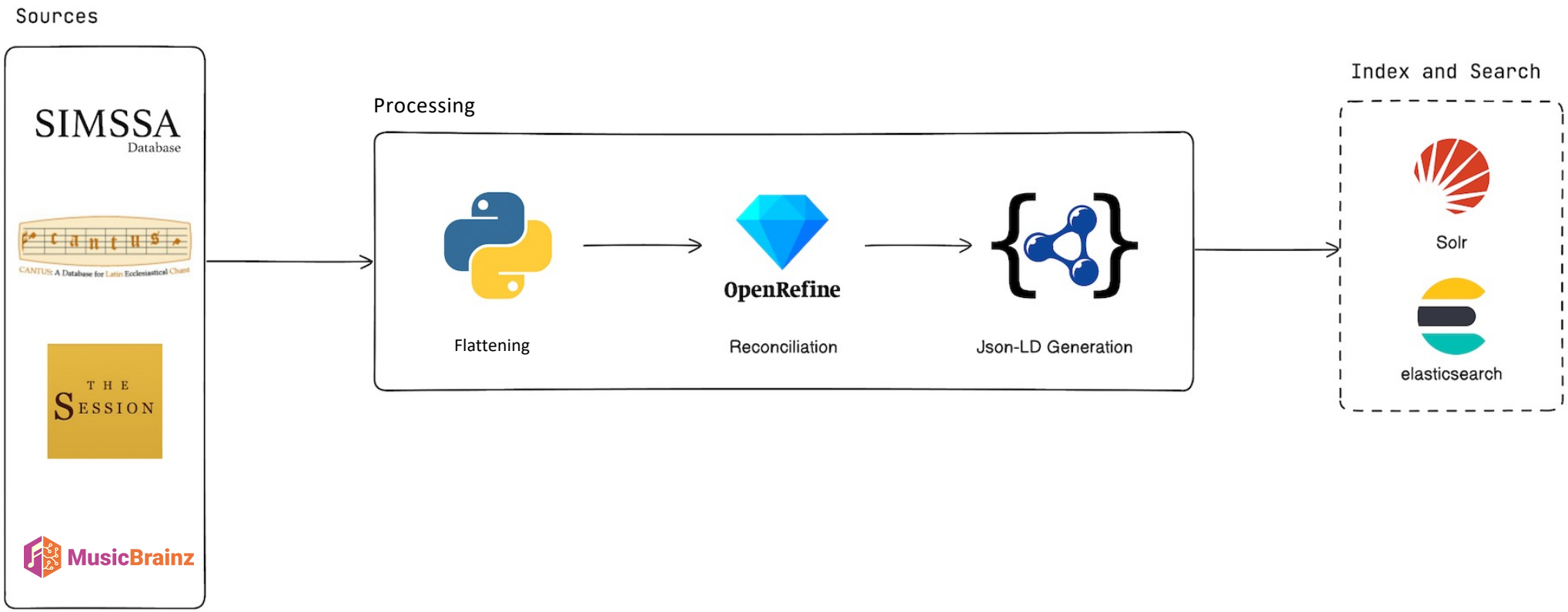
October 2023

Hong Van Pham

Jacob deGroot-Maggetti

LinkedMusic

Current Pipeline and Progress



Processing

- Foundation for the Data lake
- First step of the process

Gather data

Getting raw data from databases (CSVs, JSON, Data dump, ...)

Flatten

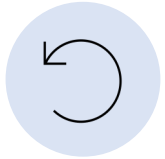
Using Python to flatten and restructure the database to a single table

Reconcile

Using OpenRefine, streamlines data reconciliation into Wikidata for improved data integrity and compatibility.

Construct JSON-LD

Create a valid JSON-LD representation of the dataset to improve discoverability and integration with Linked Data movement



Automate

Repeat the process to acquire the most recent versions, incorporate updates from databases.

Flattening process

- Flattening
- Central Entity

Flattening

- Converting relational database into nested data
- Why flatten?
  - JSON-LD formation -> Schema-free, flexible
  - Facilitate Reconciliation

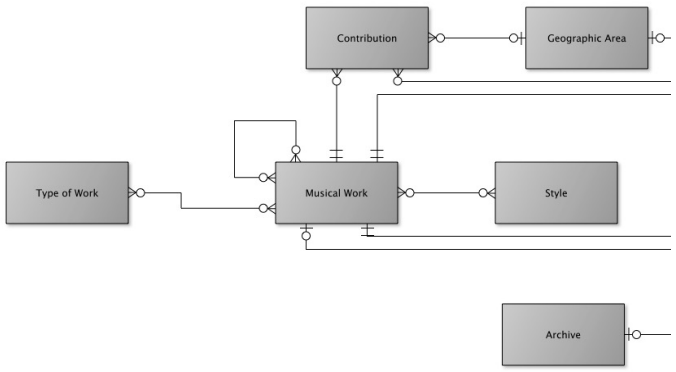
Form JSON-LD


```
{       "@id": "mw:13",
        "@type": "wd:Q2188189",
        "@context": {...},

        "database": "simssadb:",
        "musical_work_variant_titles": "['Giamai non veder gli occhi']",
        "composer": {
            "@id": "wd:Q2920493",
            "name": "Bernardo Pisano"
        },
        "genre_style": {
            "@id": "wd:Q4692",
            "name": "Renaissance"
        },
        "genre_type": {
            "@id": "wd:Q193217",
            "name": "Madrigal"
        },


        "files": [
            {
                "@type": "simssadb_file",
                "@id": "https://db.simssa.ca/files/49",
                "file_format": {
                    "@id": "wd:Q2115",
                    "name": "xml"
                },
                "Last_Pitch_Class": "[0.0]"
            },
```


reconcile



Flatten

work_id	titles	composer	genre_style	file_format	url_to_file
1	['Quanto piu desia	Bernado Pisano	Renaissance	xml	https://db.simssa.ca/files/1
1	['Quanto piu desia	Bernado Pisano	Renaissance	midi	https://db.simssa.ca/files/2
1	['Quanto piu desia	Bernado Pisano	Renaissance	pdf	https://db.simssa.ca/files/3
1	['Quanto piu desia	Bernado Pisano	Renaissance	sibelius	https://db.simssa.ca/files/4
2	['Si è debile el filo	Bernado Pisano	Renaissance	xml	https://db.simssa.ca/files/5
2	['Si è debile el filo	Bernado Pisano	Renaissance	midi	https://db.simssa.ca/files/6
2	['Si è debile el filo	Bernado Pisano	Renaissance	pdf	https://db.simssa.ca/files/7
2	['Si è debile el filo	Bernado Pisano	Renaissance	sibelius	https://db.simssa.ca/files/8
2	['Si è debile el filo	Francesco Petrarca	Renaissance	xml	https://db.simssa.ca/files/5
2	['Si è debile el filo	Francesco Petrarca	Renaissance	midi	https://db.simssa.ca/files/6
2	['Si è debile el filo	Francesco Petrarca	Renaissance	pdf	https://db.simssa.ca/files/7
2	['Si è debile el filo	Francesco Petrarca	Renaissance	sibelius	https://db.simssa.ca/files/8

Choosing the central entity

- Criteria:
  1. Data Content:
     - Nature of the data and its real-world relationships.
  2. Database Specification:
     - Table relationships and central merging identifiers

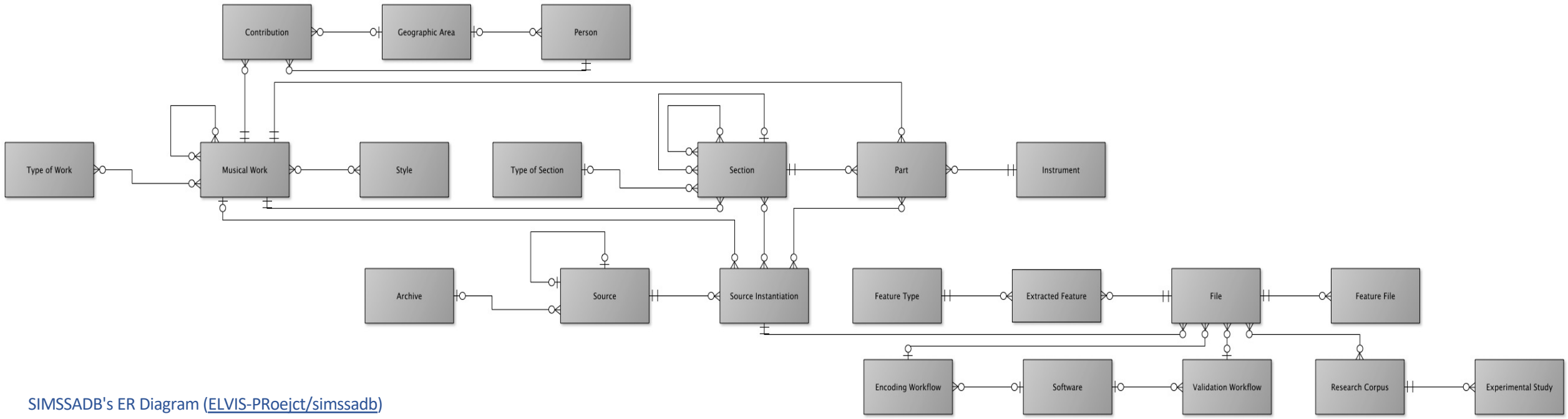
- Currently 1 central entity/database. Possible increase in the future

SIMSSA
Database

MusicBrainz

Central Entity of each Database

SIMSSA Database



SIMSSADB's ER Diagram (ELVIS-PRoejct/simssadb)

SIMSSA Database

Central entity: Musical Work

- Simssadb stores different file formats of the same identical work
- Future expansion: Section as another central entity

work_id	titles	composer	genre_style	file_format	url_to_file
1	['Quanto piu desia	Bernado Pisano	Renaissance	xml	https://db.simssa.ca/files/1
1	['Quanto piu desia	Bernado Pisano	Renaissance	midi	https://db.simssa.ca/files/2
1	['Quanto piu desia	Bernado Pisano	Renaissance	pdf	https://db.simssa.ca/files/3
1	['Quanto piu desia	Bernado Pisano	Renaissance	sibelius	https://db.simssa.ca/files/4
2	['Si è debile el filo	Bernado Pisano	Renaissance	xml	https://db.simssa.ca/files/5
2	['Si è debile el filo	Bernado Pisano	Renaissance	midi	https://db.simssa.ca/files/6
2	['Si è debile el filo	Bernado Pisano	Renaissance	pdf	https://db.simssa.ca/files/7
2	['Si è debile el filo	Bernado Pisano	Renaissance	sibelius	https://db.simssa.ca/files/8
2	['Si è debile el filo	Francesco Petrarca	Renaissance	xml	https://db.simssa.ca/files/5
2	['Si è debile el filo	Francesco Petrarca	Renaissance	midi	https://db.simssa.ca/files/6
2	['Si è debile el filo	Francesco Petrarca	Renaissance	pdf	https://db.simssa.ca/files/7
2	['Si è debile el filo	Francesco Petrarca	Renaissance	sibelius	https://db.simssa.ca/files/8


```
{       "@id": "mw:13",
        "@type": "wd:Q2188189",
        "@context": {...},

        "database": "simssadb:",
        "musical_work_variant_titles": "['Giamai non veder gli occhi']",
        "composer": {
            "@id": "wd:Q2920493",
            "name": "Bernardo Pisano"
        },
        "genre_style": {
            "@id": "wd:Q4692",
            "name": "Renaissance"
        },
        "genre_type": {
            "@id": "wd:Q193217",
            "name": "Madrigal"
        },


        "files": [
            {
                "@type": "simssadb_file",
                "@id": "https://db.simssa.ca/files/49",
                "file_format": {
                    "@id": "wd:Q2115",
                    "name": "xml"
                },
                "Last_Pitch_Class": "[0.0]"
            },
```

The Session

Central entity: tunes

- Based on database structure

tune_id	name	setting_id	type	meter	mode	username
5478	'S Iomadh Rud A Chunnaic Mi	5478	reel	4/4	Gmajor	Andy F
5478	'S Iomadh Rud A Chunnaic Mi	11429	reel	4/4	Dmajor	malcombpiper
15326	S Ann An Ìle	28560	strathspey	4/4	Gmajor	danninagh
15326	S Ann An Ìle	28582	strathspey	4/4	Gmajor	DonaldK
14625	'S Daor An Tabac	26955	reel	4/4	Bminor	Charles Mackenzie
7078	10th Bat Crossing Rhine	7078	jig	6/8	Amixolydian	gaitazampogna_32
7078	10th Bat Crossing Rhine	18649	jig	6/8	Amixolydian	ceolachan
7078	10th Bat Crossing Rhine	18650	jig	6/10	Dmajor	ceolachan


```
{
    "@id": "tunes:1",
    "@type": "wd:Q2188189",
    "@context": {...}
    "popularity_tunebooks": 5304,
    "tunes_name": "Cooley's",
    "tunes_type": {
         "@id": "https://thesession.org/tunes/search?type=reel",
         "name": "reel"
    },
    "recordings": [
       {
         "@id": "thesession:recordings/3536",
         "@type": "wd:Q3302947",
         "artist": "Accord\u00e9onistes Du Qu\u00e9bec",
         "recording": "Dans Tous Les Cantons",
         "track": 23
        },
      ....
    ],
    "alias": "['Cooleys', 'Joe Cooley']",
    "settings": [
       {
         "tunes_type": "reel",
         "@id": "https://thesession.org/tunes/1#setting1",
         "@type": "wd:Q113899068",
         "meter": "4/4",
         "mode": "Edorian",
         "date": "2001-05-14T18:45:18",
         "username": "Jeremy"
        },
      .....
     ]
}
```

Cantus Database

Central entity: Chant

- While CantusDB chants share similarities under the same CantusID, the database emphasizes their distinctions.


```
{
        "incipit": "Accepit autem omnes timor et ",
        "finalis": "D",
        "absolute_url": "https://cantusdatabase.org/chant/561923",
        "@context": {...},
        "@id": "chant:561923",
        "@type": "wd:Q23072435",
        "database": "cantusdb:",
        "composer": {
            "@id": "wd:Q4233718",
            "name": "Anonymous"
        },
        "genre": {
            "@id": "wd:Q582093",
            "name": "Antiphon"
        },
        "mode_name": {
            "@id": "wd:Q960729",
            "name": "dorian"
        },
        "source": "https://cantusdatabase.org/source/123756",
        "cantus_id": "https://cantusindex.org/id/001216"
    }
```

Data Export from CantusDB to LinkedMusic Data Lake

- Initial export of 50 chants
- To begin, focusing on metadata (i.e., things we can reconcile with WikiData)
- Eventually, data (text, melodic transcriptions, etc.) will be used too

Contents of initial sample

- ID
- Incipit
- Genre (e.g., "Responsory verse", "Antiphon")
- Cantus ID
- Finalis (e.g., "D", "E", ...)
- Mode (e.g., "1", "2", ..., "8", "4T", "7?")
- *Mode name (e.g., "dorian", "hypodorian", ..., "hypomixolydian")*
- *Composer (always "Anonymous")*
- *Absolute URL, Source Link*

MusicBrainz

- MusicBrainz provides JSON-LD data
- Some items are reconciled with Wikidata

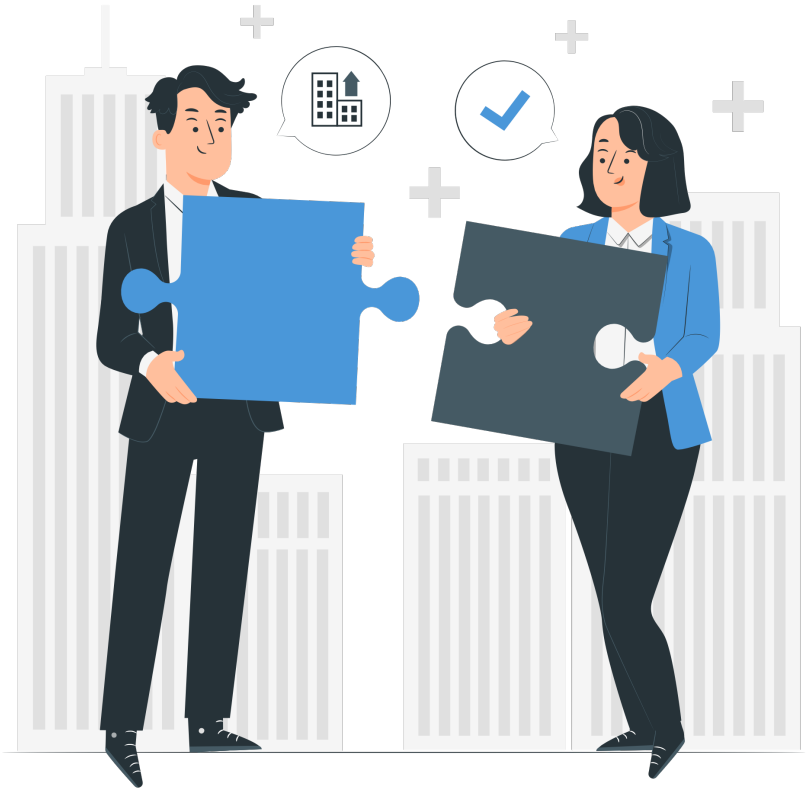
- Limitation with OpenRefine: Doesn't handle nested data

-> Use **Reconciliation Service API** (bypass OpenRefine)

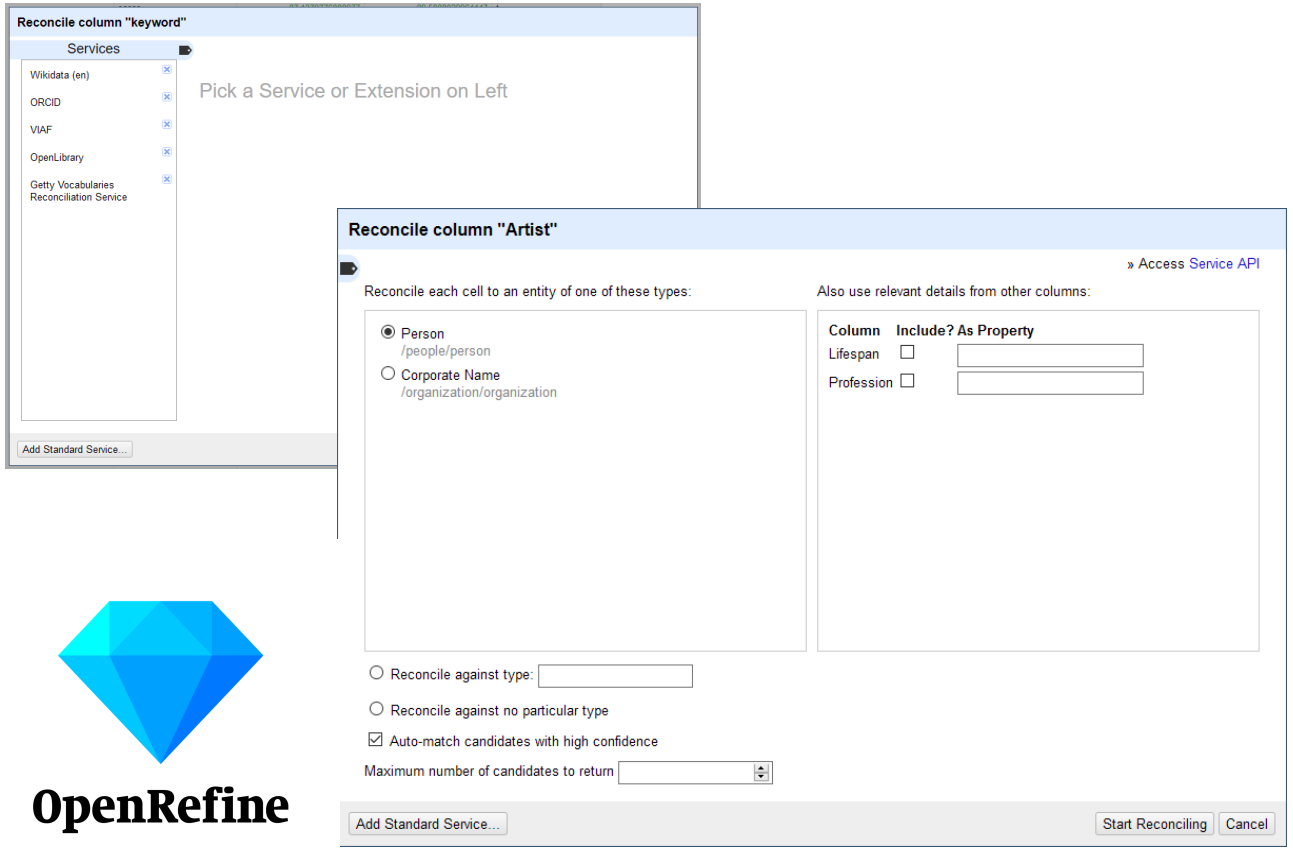

```
{
    "@id": "http://musicbrainz.org/release/20027ad4-6667-49b6-bd07-5ac12
    "@type": "MusicRelease",
    "@context": [...]
    "recordLabel": {
        "@id": "http://musicbrainz.org/label/d476b74d-670d-44ab-be5c-2c1
        "name": "Philo",
        "@type": "MusicLabel"
    },
    "name": "McGreevy & Cooley",
    "image": [...],
    "releaseOf": {
        "creditedTo": "Johnny McGreevy & S\u00e9amus Cooley",
        "albumProductionType": "http://schema.org/StudioAlbum",
        "@type": "MusicAlbum",
      ....
    },
    "track": [
      {
        "name": "The Broken Pledge / Julia Delaney (reels)",
        "@id": "http://musicbrainz.org/recording/d506fa4b-733e-4864-8b41
        "trackNumber": "1.1",
        "@type": "MusicRecording"
      },
      {
        "name": "Se\u00e1n sa Ceo / Michael Preston's (reels)",
        "@id": "http://musicbrainz.org/recording/b59c6d66-e6c1-4902-a673
        "@type": "MusicRecording",
        "trackNumber": "1.2"
      },
     ....
    ],
    ....
}
```

Reconciliation process

- Value Reconciliation
- Properties Reconciliation



Value Reconciliation process



OpenRefine

- Matching data with Wikidata's items (Qs) to improve searchability and quality.
- OpenRefine offers reconciliation capabilities for Wikidata IDs based on Reconciliation Service API.
- Functions:
  - Reconciliation based on data types.
  - Using multiple columns to facilitate reconciliation.

Demo

Value Reconciliation process

Demo

work_id	musical_work_variant_titles	composer	composer_@id	file_format	file_format_@id
1	['Quanto più desiar sento'l mio (	Bernardo Pisano	Q2920493	xml	Q2115
1	['Quanto più desiar sento'l mio (	Bernardo Pisano	Q2920493	xml	Q2115
1	['Quanto più desiar sento'l mio (	Bernardo Pisano	Q2920493	midi	Q10610388
1	['Quanto più desiar sento'l mio (	Bernardo Pisano	Q2920493	xml	Q2115
3	['Si è debile el filo a cui s'attiene	Bernardo Pisano	Q2920493	xml	Q2115
3	['Si è debile el filo a cui s'attiene	Bernardo Pisano	Q2920493	pdf	Q42332
3	['Si è debile el filo a cui s'attiene	Bernardo Pisano	Q2920493	xml	Q2115
3	['Si è debile el filo a cui s'attiene	Bernardo Pisano	Q2920493	xml	Q2115
4	['Amor quando fioriva mia sperr	Sebastiano Festa	Q7442579	midi	Q10610388
10	['Amor quando fioriva mia sperr	Sebastiano Festa	Q7442580	xml	Q2115
10	['Amor quando fioriva mia sperr	Sebastiano Festa	Q7442581	xml	Q2115
10	['Amor quando fioriva mia sperr	Sebastiano Festa	Q7442582	pdf	Q42332
10	['Amor quando fioriva mia sperr	Sebastiano Festa	Q7442583	midi	Q10610388
10	['Amor quando fioriva mia sperr	Sebastiano Festa	Q7442584	xml	Q2115

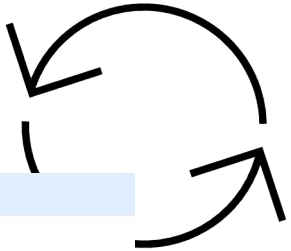
- We reconcile to obtain Wikidata IDs while keeping the original content from the database

- Currently only reconcile certain columns:
  - Of type with more items available on Wikidata
  - Have repeating values
- Example:
  - "Composer"
  - "Mode name"
  - "Location"
  - …
- Future:
  - Reconcile as many values as we can
  - Create new items on Wikidata

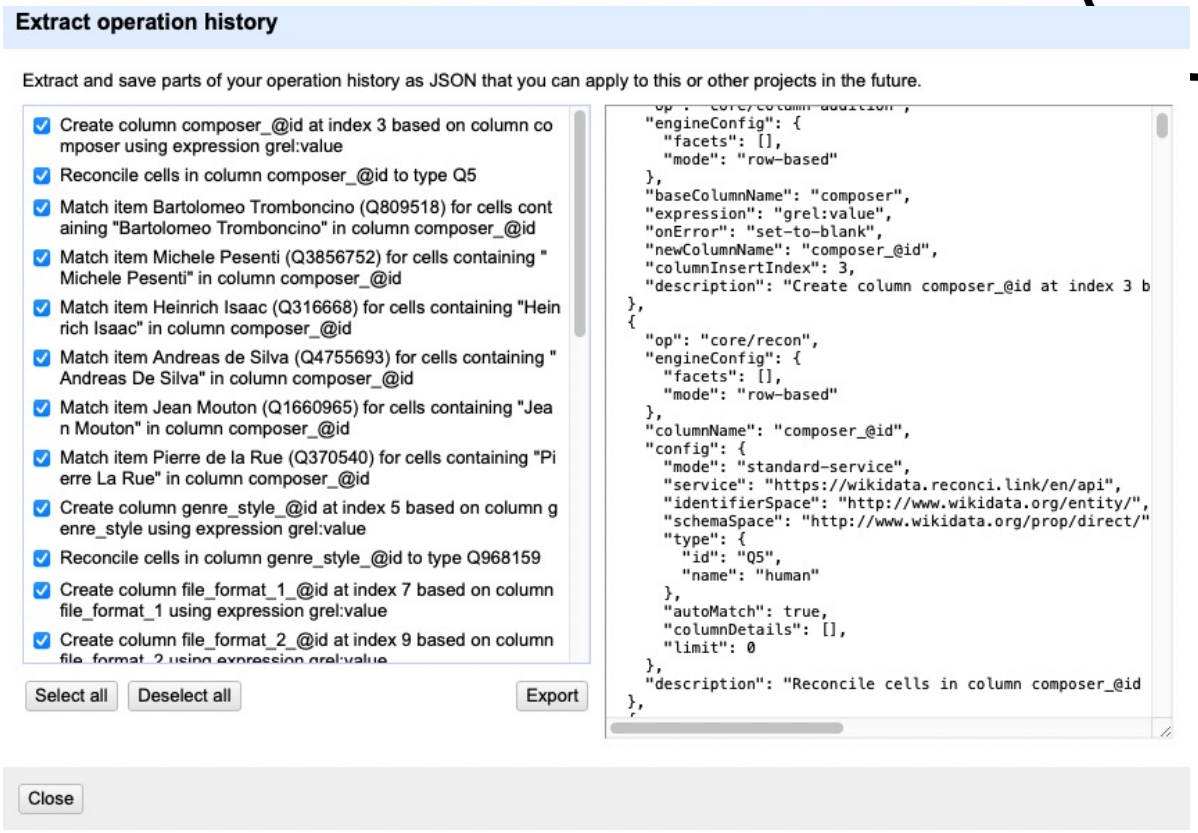
OpenRefine

Demo

Re-reconciliation



- Operations on OpenRefine can be reused
- We can Extract the operation sequence (including Reconciliation choices) and reapply them for later projects (conditions apply)



OpenRefine

Automation

- Pipelines are interrupted by OpenRefine manual reconciliation

- Automation: opencultureconsulting/**orcli**
  - OpenRefine Command line interface
  - Uses the operation history and export template

- Another option: use the Reconciliation Service API directly

- Future project: Frequent data update!

OPEN CULTURE
CONSULTING

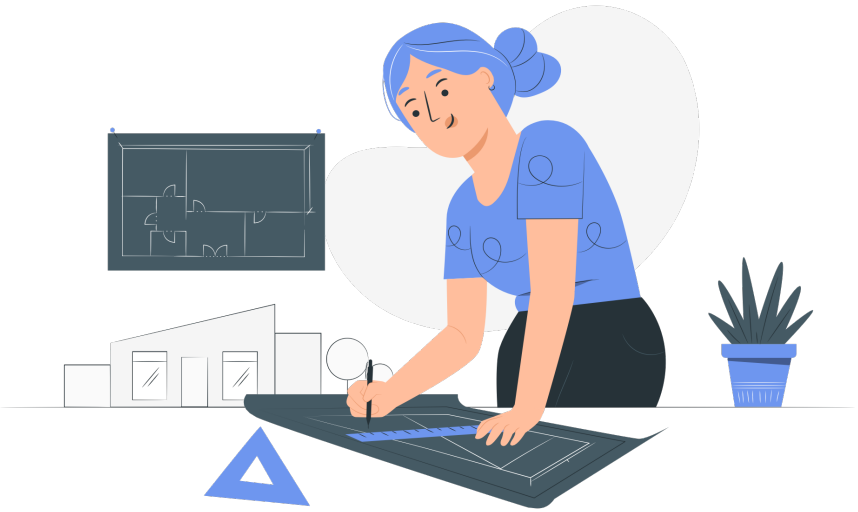
German IT Consulting for Cultural Heritage and Academic Institutions

Properties Reconciliation

- We can leave cells un-reconciled, but we can't leave properties un-reconciled!
- Done manually in the JSON-LD context
- Priority order:
  - Wikidata Properties
  - Schema.org Properties
  - Database's documentations
  - Other trusted source

JSON-LD Structure

- JSON-LD Context and Document
- Expansion



JSON-LD Context

- Provides context to make JSON readable without losing information
- Link the database properties to corresponding Wikidata Properties (Ps)

```
{
    "@context": {
        "wd": "http://www.wikidata.org/entity/",
        "wdt": "http://www.wikidata.org/prop/direct/"

        "simssadb": "https://db.simssa.ca/",
        "mw": "simssadb:musicalworks/",
        "simssadb_file": "simssadb:files/",
        "files": "simssadb_file",

        "musical_work_variant_titles": "wdt:P1476",
        "composer": "wdt:P86",
        "genre_style": "wdt:P135",
        "genre_type": "wdt:P136",
        "file_format": "wdt:P2701",
        "name": "wd:P2561",

        "database": {
            "@id": "https://schema.org/Dataset",
            "@type": "@id"
        },

    }
}
```

JSON-LD Document

- Normal JSON file representing the database, human readable
- Each document represent one instance of the central entity
- Reconciled properties store both external URI ("@id") and original value ("name")
- Include "@context" which points to a JSON-LD context
- Referenced the JSON-LD structure of MusicBrainz and Wikidata to design our own.


```
{       "@id": "mw:13",
        "@type": "wd:Q2188189",
        "@context": {...},

        "database": "simssadb:",
        "musical_work_variant_titles": "['Giamai non veder gli occhi']",
        "composer": {
            "@id": "wd:Q2920493",
            "name": "Bernardo Pisano"
        },
        "genre_style": {
            "@id": "wd:Q4692",
            "name": "Renaissance"
        },
        "genre_type": {
            "@id": "wd:Q193217",
            "name": "Madrigal"
        },


        "files": [
            {
                "@type": "simssadb_file",
                "@id": "https://db.simssa.ca/files/49",
                "file_format": {
                    "@id": "wd:Q2115",
                    "name": "xml"
                },
                "Last_Pitch_Class": "[0.0]"
            },

            {...}
        ]
```


1 JSON-LD Document of SIMSSADB, representing 1 musical work

Adding Contexts...

```
{
    "@context": {
        "wd": "http://www.wikidata.org/entity/",
        "wdt": "http://www.wikidata.org/prop/direct/"

        "simssadb": "https://db.simssa.ca/",
        "mw": "simssadb:musicalworks/",
        "simssadb_file": "simssadb:files/",
        "files": "simssadb_file",

        "musical_work_varia
        "composer": "wdt:P
        "genre_style": "wd
        "genre_type": "wdt
        "file_format": "wd
        "name": "wd:P2561",

        "database": {
            "@id": "https:/
            "@type": "@id"
        },

    }
}
```

```
{
    "@id": "mw:13",
    "@type": "wd:Q2188189",
    "@context": {...},

    "database": "simssadb:",
    "musical_work_variant_titles": "['Giamai non veder gli occhi']",
    "composer": {
        "@id": "wd:Q2920493",
        "name": "Bernardo Pisano"
    },
    "genre_style": {
        "@id": "wd:Q4692",
        "name": "Renaissance"
    },
    "genre_type": {
        "@id": "wd:Q193217",
        "name": "Madrigal"
    },


    "files": [
        {
            "@type": "simssadb_file",
            "@id": "https://db.simssa.ca/files/49",
            "file_format": {
                "@id": "wd:Q2115",
                "name": "xml"
            },
            "Last_Pitch_Class": "[0.0]"
        },

        {...}
    ]
```

Expanded

```
{
  "@id": "https://db.simssa.ca/musicalworks/13",
  "@type": "http://www.wikidata.org/entity/Q2188189",
  "http://www.wikidata.org/prop/direct/1476": "['Giamai non veder gli occhi']"
  "http://www.wikidata.org/prop/direct/P135": {
    "@id": "http://www.wikidata.org/entity/Q4692",
    "http://www.wikidata.org/entity/P2561": "Renaissance"
  },
  "http://www.wikidata.org/prop/direct/P136": {
    "@id": "http://www.wikidata.org/entity/Q193217",
    "http://www.wikidata.org/entity/P2561": "Madrigal"
  },
  "http://www.wikidata.org/prop/direct/P86": {
    "@id": "http://www.wikidata.org/entity/Q2920493",
    "http://www.wikidata.org/entity/P2561": "Bernardo Pisano"
  },
  "https://db.simssa.ca/files/": {
    "@id": "https://db.simssa.ca/files/49",
    "@type": "https://db.simssa.ca/files/",
    "http://www.wikidata.org/prop/direct/P2701": {
      "@id": "http://www.wikidata.org/entity/Q2115",
      "http://www.wikidata.org/entity/P2561": "xml"
    }
  },
  "https://schema.org/Dataset": {
    "@id": "https://db.simssa.ca/"
  }
}
```

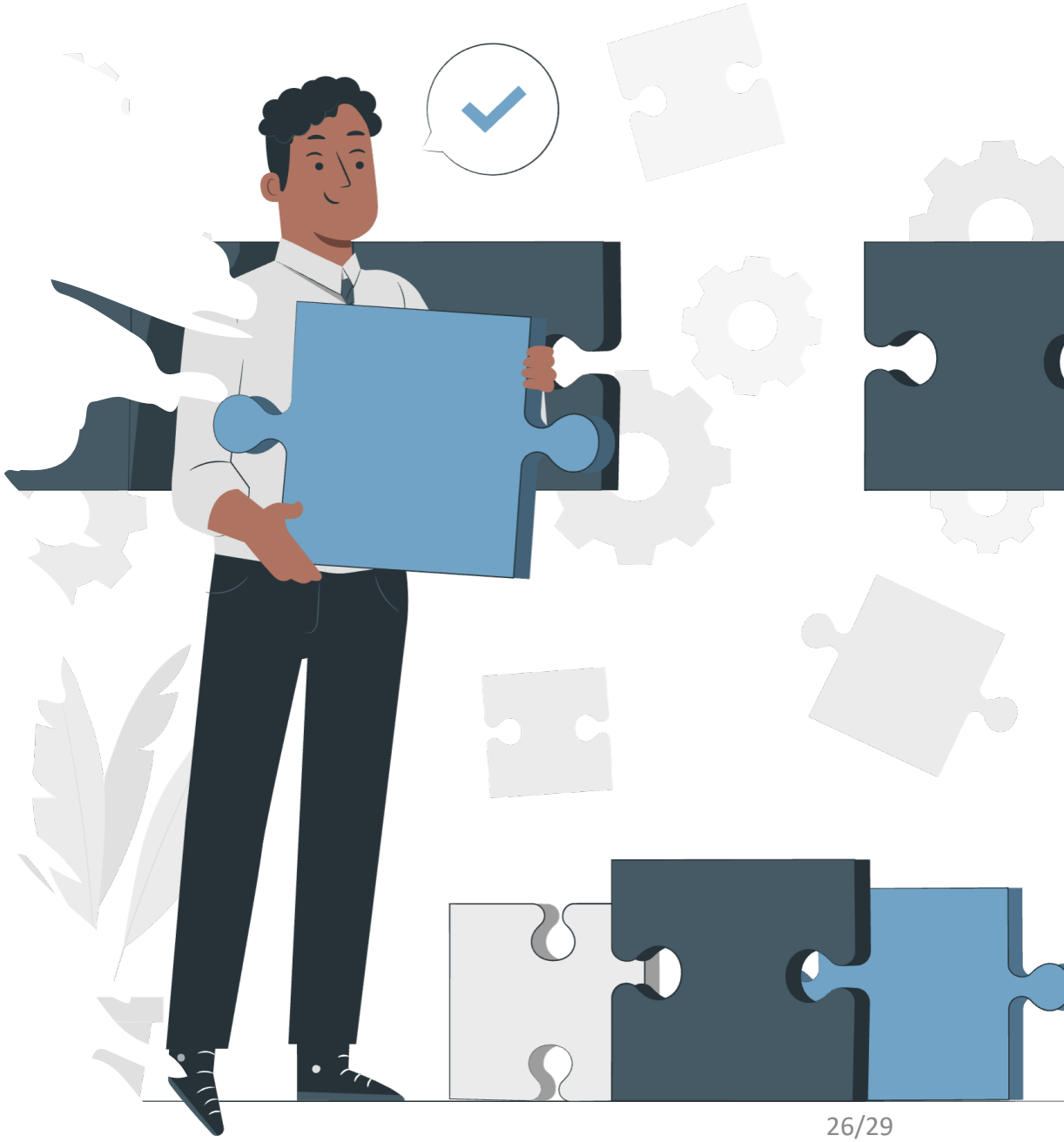
Difficulties

- Difficulties in mapping properties with Wikidata Properties
  - How to handle Non-existing/incomplete match?
  - Who decides?
  - Creating properties (Ps) on Wikidata is difficult.

- OpenRefine doesn't handle nested structure very well
  - Ex: Musicbrainz provides JSON-LD but export from OpenRefine can't export the structure
  - Implement directly the Reconciliation Service API

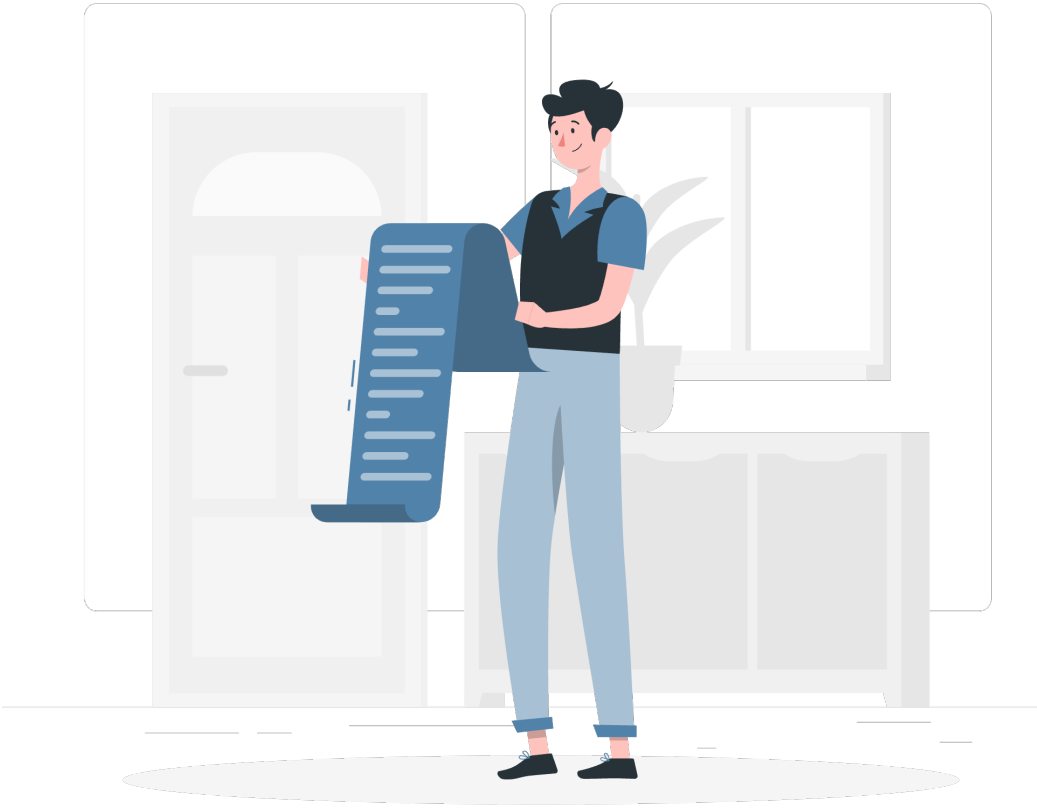
Future Objectives

- Expand the project to other databases
- Capture full database architecture
- Automate and schedule pipelines



Summarize

- Overall Process
- Flattening and Central entity
- Values and Properties Reconciliation & Re-reconciliation
- JSON-LD and context structure
- Difficulties and Future objectives



Thank you for your attention

van.pham2@mail.mcgill.ca
jacob.degroot-maggetti@mail.mcgill.ca

DDMAL DISTRIBUTED DIGITAL MUSIC ARCHIVES & LIBRARIES LAB

LinkedMusic

McGill

Question