The workflow shown in figure 1. The final dataset can be gathered with the help of automatic, semi-automatic or manual processes.

Diagram

Description automatically generated

Figure 1. Workflow and metadata of DSI members publications

**Step1/Milestone-1: Automatic/Semi-Automatic**

We should use Cecile’s excel worksheet to extract individual profile and then we should locate that profile in DBLP dataset. From DBLP dataset, we can extract all the publications of the individual. DBLP generates XML on their site, as you can seen in the below snippet for “Denise McGrath”.

Graphical user interface, text, application, email

Description automatically generated

Thus, we will be able to extract all the publications and their metadata as depicted in Figure 1. DBLP identify authors with ids, thus you will be able to find unique authors. We should keep two distinct files, i.e., For keeping record of authors and co-authors 2. The publication details.

1. **For authors and co-authors** – it should be like the Cecile’s file. Here is the snapshot of its structure

Graphical user interface, table, Excel

Description automatically generated

1. **The publication details**

This should be another file/excel-sheet where publication datils should be kept. For the time being, we can save the data proposed in Figure 1. The file in 1 and 2 will be linked up by the publication Id.

In DBLP, we can find author ORCID and from <https://orcid.org> we can extract the authors affiliation details. Here is the snippet

Graphical user interface, text, application, email

Description automatically generated

Following structure, we can use for our dataset

Author/Co-Author File Structure:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| PublicationID | AuthorID | AuthorName | Author Affiliations | Rank (optional) |

Publication File Structure:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| PublicationID | Title | PublicationDate | Venue | URL | Publication Type |  |

Till this, the process can be completed automatically/semi-automatically. And it should not take much longer time.

**Step2: Extracting Publication Additional Details – Manual, Semi-Automatic, o Automatic**

This is the next stage of data gathering, we can think of extracting abstract information of the publications. As explained earlier, we will have publication access URL, we can think of extracting Abstract manually or automatically.