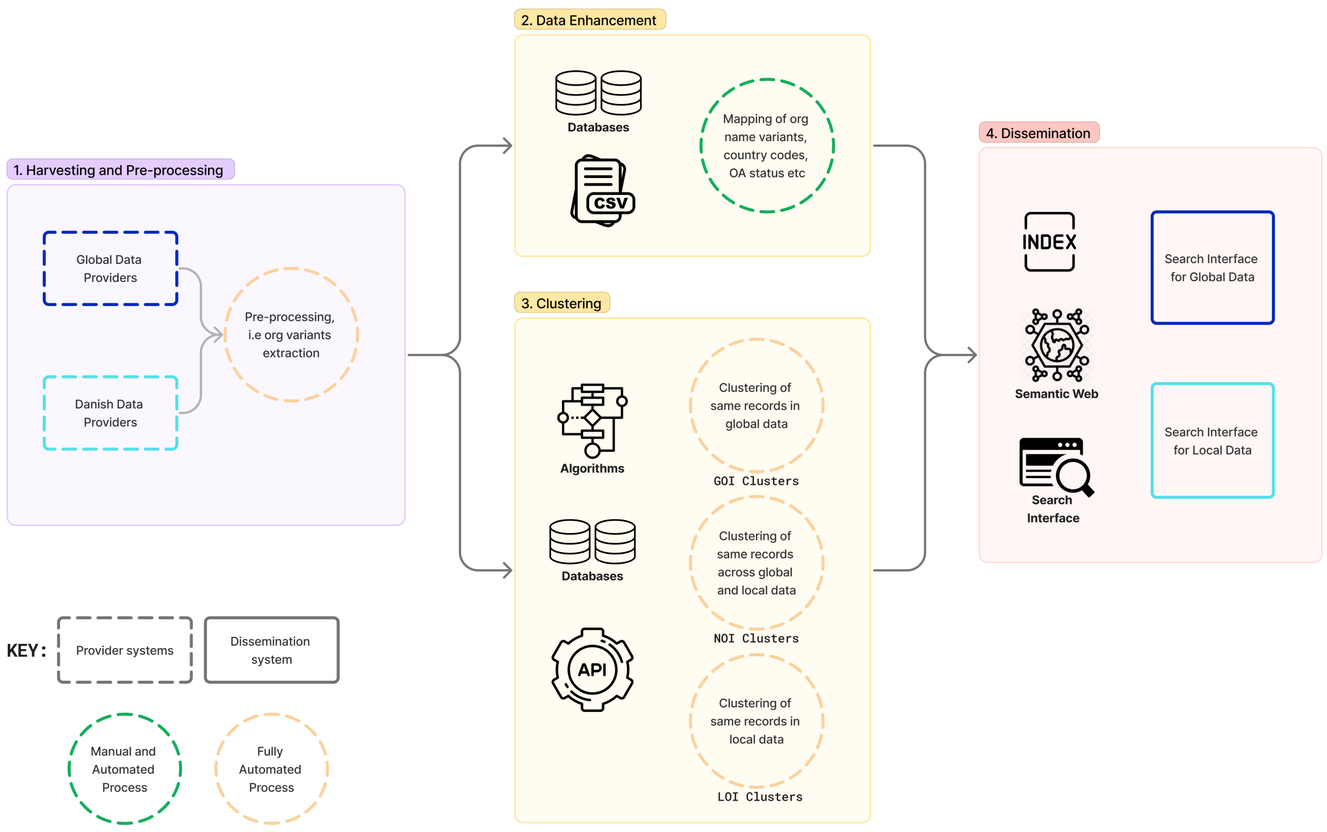
Architecture and Data Flow Diagrams

How everything works together



High level data flow

1. Harvesting and pre-processing: The data from different commercial and local data providers respectively is gathered
and stored.

2. Data Enhancement: The data is augmented and enhanced with additional or improved data elements, in collaboration between the main pipelines and the NORA team's data analysts.

3. Data Consolidation: The data is clustered to identify and link identical publication records from different providers.

4. Dissemination: The data is made available through a web/search interface and analytical overviews.

Architecture in more Detail

There are two main pipelines the **Global** pipeline and the **Local** pipeline. The main process of data enhancement and consolidation are common, with the main steps being:

- Extracting and counting the list of unique affiliation IDs and organisation name variants to feed into the affiliation mappings
- Clustering of global data (GOI) + clustering of local data (LOI) ⟶ clustering and matching between global and local data (NOI clusters)

The harvesting and the presentation are handled differently due to the many ways of accessing the data from the providers, the different data quality and data formats.