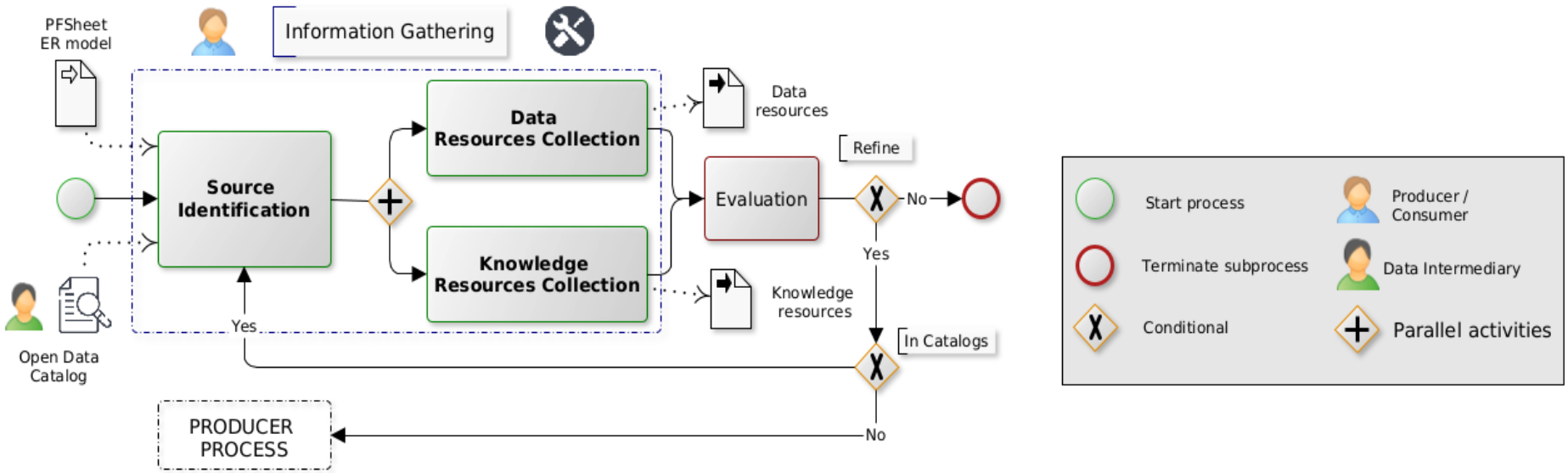
# Part 4.3
# Phase 2 - Information Gathering

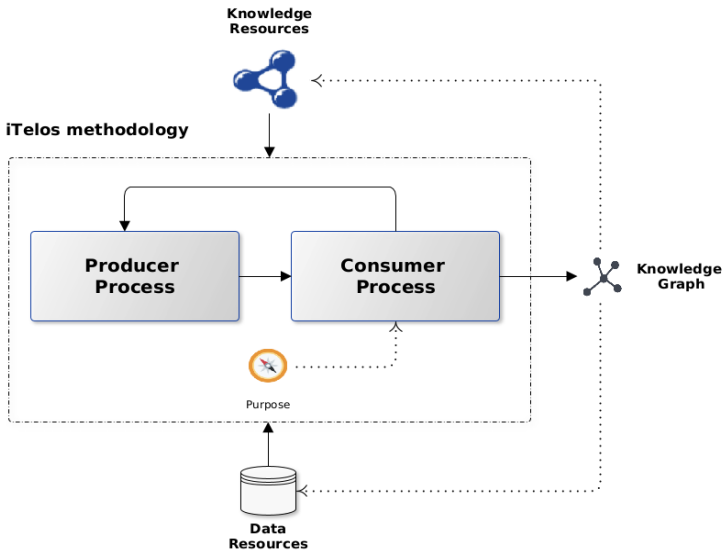
# Phase 2 - Information Gathering




- **Input**: Purpose Formalization sheet, ER model, Source list.
- **Objective**: collecting the resource, to be processed, to build the final KG(s), thus satisfying the formalized purpose.
- **Output**: Data and Knowledge Resources (Formal, Semi-Formal or Informal).

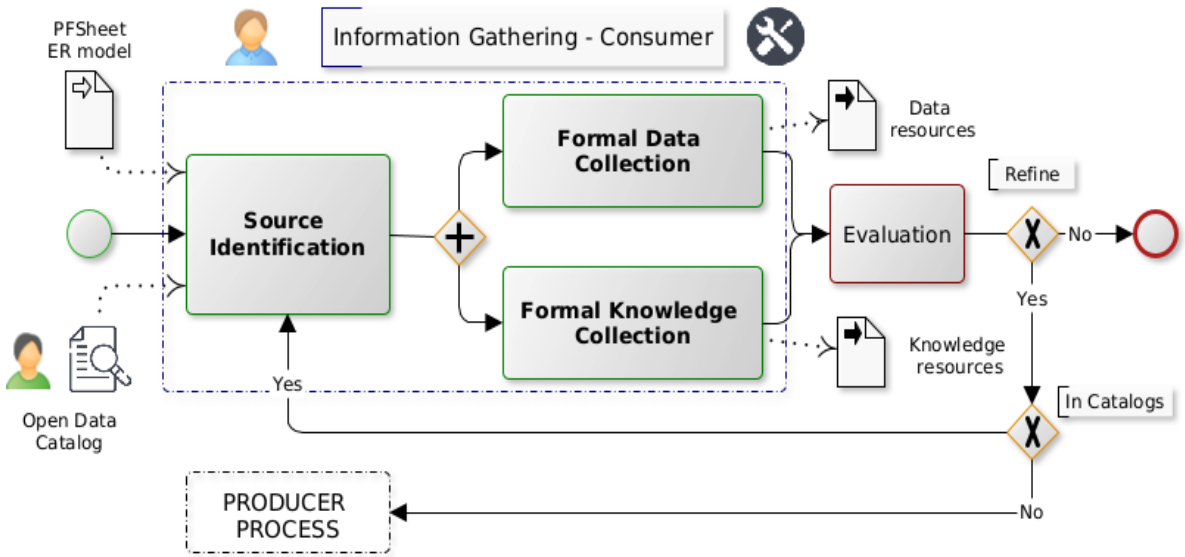
# Phase 2 - Information Gathering - Producer & Consumer

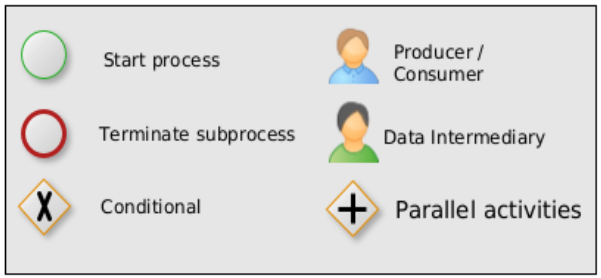
- This is the phase where (more then the others) the distinction between **Data Producer** and **Data Consumer** appears.
- Consumer side, the objective of this phase is to collect **formal resources** to be composed, with the objective of building the final KG.
- If the formal resources available are not sufficient, the iTelos methodology **leads to the execution of the process at Producer side**, where the required formal resources are produced.



# Phase 2 - Information Gathering - Consumer

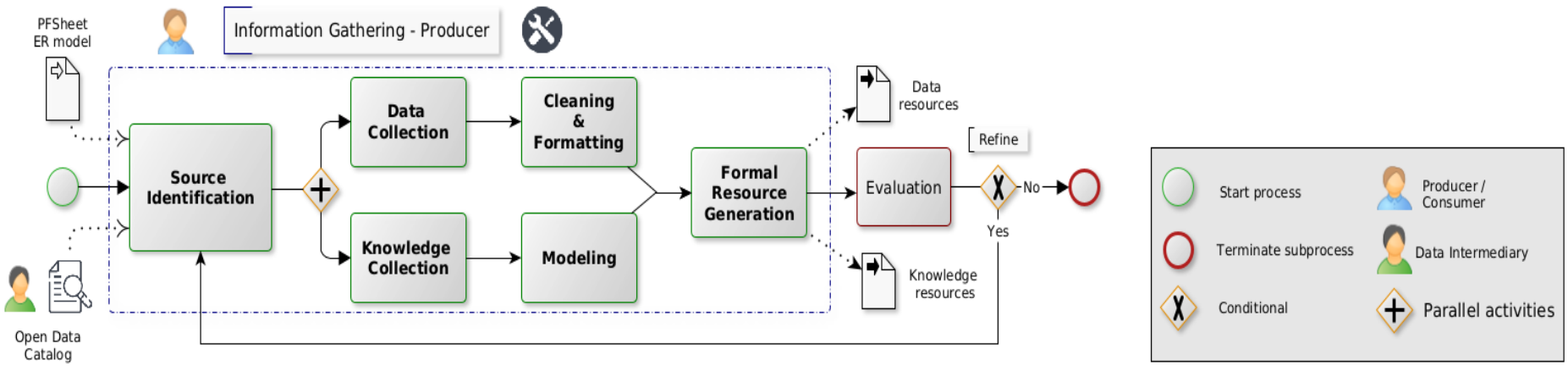
- At **Consumer** side, the data collected is already formalized (high quality and interoperability).
- Therefore the objective of this phase is to **identify** the sources and **collect** the resources from them.



# Phase 2 - Information Gathering - Producer

- At **Producer** side the data collected is informal.
- The heterogeneity on such a data has to be handled at each level (source, format, structure and meaning). For this reason more activities are considered at Producer side.[37]




[37] We will describe the producer side activity set, which includes the Consumer side activities as well.

# Information Gathering - Source Identification

- The first activity of the current phase, aims at **identifying and accessing the sources of information** provided as input, and, eventually discovering other sources (if those in input are not sufficient).

- Depending by the type of process executed, this activity considers different kind of information sources:

  - **Consumer process**: the sources considered are Catalogs where formal resources (produced by an iTelos producer process) are distributed (i.e. DataScientia catalogs).

  - **Producer process**: the sources considered can be different and distributing different type of data, with a less or more quality.
    - Informal data catalogs;
    - Web pages;
    - Databases;
    - ....

# Information Gathering - Resource Collection

- **Consumer process**: the collection of resources, both knowledge and data, form formal catalogs, requires less effort respect the resource collection at producer side.
  - **Clear policies of data distribution**. Direct distribution/download, or distribution on demand through request to the data owner.
  - **High quality metadata**, describing the available resources. This increase the Findability ad Accessibility of the resources distributed.
- **Producer process**: the data is not always distributed, sometimes it is only **published** or **visualized** online. This means that different approaches for data collection have to be considered:
  - asking for datasets directly to owners;
  - accessing data through automatic or semi automatic portals (catalogs);
  - scraping data from sources (this usually requires scraping libraries customization);
  - producing our own data (data collection apps and tools [iLog]).

## Information Gathering - Resource Collection

- Collecting data, in general, aims to achieve the following two results:
  - Increase the number of **entities** and/or entity types.
  - Increase the number of **entity attributes** and/or entity type properties.

- Are the resources collected covering your list of CQs ?
  - **yes** - let's proceed on.
  - **no** - go back to source identification.

# Information Gathering - Data Cleaning

- The cleaning activity aims to **remove "noise"** from the set of resources collected.
  - This activity is **mainly considered for the Producer process**. Nevertheless, it could happen that at Consumer side some informal resources have to be cleaned.

- "Noise" is intended to be:
  - entire **datasets** without any information to be considered to satisfy the Purpose;
    - (it happens often collecting automatically or receiving huge amount of data)
  - **etypes and/or entities**, within single datasets, with no relevance for the Purpose;
  - **properties and/or attribute**, within single datasets, with no relevance for the Purpose.

# Information Gathering - Data Formatting

- Now the set of resources (both knowledge and data) has been finalized.
- The Formatting activity aims to:
  - **align the differ formats** present in the heterogeneous resource set (datasets formats and data values formats (data types));
  - **anonymise the data** collected; required only if sensible information (like personal data) are included in the datasets collected.
- This activity is mainly executed at Producer side.

**Note**: the format alignment over common standards (CSV, XML, TSV, JSON, RDF and OWL) is strongly required, mainly for two reasons:

- Reusability.
- Compliance with iTelos process.

# Information Gathering - Knowledge Modeling

- In order to produce Formal resources the dataset collected, cleaned and formatted have to be **associated to a schema** representing the information their are carrying.
- This activity is mainly executed at **Producer side**.
- Such a schema, **for each single datasets**, has to be formally defined (in RDF-OWL formats). How to define a schema for the single dataset ?
  - The dataset structure is **self-explanatory**, thus reducing the modeling effort (no conceptual integration between two or more datasets).
  - The dataset's **information needs to be interpreted**, thus a point of view is required for such an interpretation.
    - Which one ? **The Purpose**.
- Therefore, also at Producer side the Purpose plays a crucial role, not for the integration/composition of resources, but for their single interpretation.
  - **There is no data without a purpose!**

# Information Gathering - Formal Resource Generation

- With the objective of building formal resources in KG-based form, the Formal Resource Generation activity takes in input:
  - (data layer) The datasets collected, cleaned and formatted.
  - (knowledge layer) The schemas produced (extracted) for each datasets.
- This activity produces in output, **for each pair composed by a dataset and its relative schema**, a single object representing a **formal KG**.
- To achieve this result a specific tool is offered by iTelos (Karmalinker) used to map each dataset over its own schema, thus merging data and knowledge layer of a KG. [38]
- This activity is mainly executed at **Producer side**.

[38] This tool is exploited also in the last iTleos phase when the final KG is built

# Information Gathering - Tools and Standards

- **Source identification**: the activity requires only the web navigation, looking for catalogs and other information sources.
- **Resource collection, Cleaning & Formatting**:
  - **Consumer**: catalogs resource distribution services (download, and download request).
  - **Producer**: data scraping and management libs:
    - **Dataframe**: Pandas
    - **Calcs**: NumPy
    - **Plotting**: Matplotlib, graphviz
    - **REST API**: Requests
    - **Scraping**: Scrapy, BeautifulSoup4
    - **Dates**: Dateparser, Arrow
    - **Geospatial**: GeoPandas, geopy, geoplot (Cartopy)

## Information Gathering - Tools and Standards

- **Modeling**: the modeling of schema for each dataset collected, is performed using the **Protègè** tool, by creating **RDF-OWL** schema files.
- **Formal resource generation**: the merging between knowledge and data layers for each pair dataset-schema collected, is performed with a specific data mapping tool called **Karmalinker**, producing **RDF** files in output.

## Information Gathering - Practical Implementation

- The practical implementation of the current phase includes:
  - The **execution of all the above described activities**, by using the tools and standards mentioned.
  - The **upload on the project github repository**, of the resources collected and formalized, during the different phase's activities.
- **Dedicated demo lecture!**