

KNOWDIVE



KGE - Knowledge Graph Engineering

iTelos Methodology

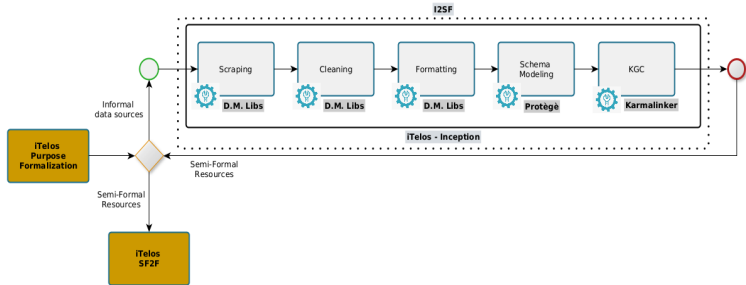
Phase 1 - Inception

Fausto Giunchiglia

Contents

- 1 Phase structure**
- 2 Inception - Input
- 3 Inception - Activities
- 4 Inception - In practice

Inception



- **Input:** a set of data sources identified previously, plus the initial user's Purpose.
- **Output:** a set of semi-formal resources, created from the informal resources extracted by the data sources in input.
- **Objective:** to extract, clean, format and model the informal resources required to satisfy the Purpose, collected from the input data sources.

Contents

- 1 Phase structure
- 2 Inception - Input**
- 3 Inception - Activities
- 4 Inception - In practice

Inception - Input

- **The Purpose:** the initial purpose is always considered as input for each iTelos phase.
 - It allows to make the right decisions over the issues to be addressed within each phase.
 - In the Inception phase, the Purpose supports the data management activities, by **clarifying the relevance of data**, and data values, to achieve the desired output.
- **Data source list:** the list of (data and knowledge) sources identified in the previous phase.

Contents

- 1 Phase structure
- 2 Inception - Input
- 3 Inception - Activities**
- 4 Inception - In practice

Inception - Activities

- The Inception phase aims to collect and semi-formalize the resources (knowledge and data) used to build the final KG.
- Due to the **heterogeneity of the sources**, provided in input, the execution of this phase can be different source by source, or even dataset by dataset.
 - different collection procedures;
 - different cleaning and formatting activities;
 - different schemas to be modeled for each resource collected;
- As a consequence, an **iterative execution** over the source list is considered for this phase.

Inception - Scraping (collection)

- The first activity of Inception phase aims to collect the required resources.
- Collecting data can be done in several ways:
 - 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]).

Inception - Scraping (collection)

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

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

Inception - Cleaning

- The cleaning activity aims to **remove "noise"** from the set of resources collected.
- "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)
 - **entities**, within single datasets, with no relevance for the Purpose;
 - entity **attributes and entity attribute's values**, within single datasets, with no relevance for the Purpose.

Inception - Formatting

- Now the set of resources (as well as entities and relative entity attributes) has been finalized.
- The Formatting activity aims to:
 - **align the differ formats** present in the heterogeneous resource set (datasets formats and data values formats);
 - **anonymize the data** collected; required only if sensible information (like personal data) are included in the datasets collected).

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

- Reusability.
- Compliance with iTelos activities.

Inception - Schema modeling

- In order to produce Semi-Formal resources the dataset collected, cleaned and formatted have to be **associated to a schema** representing the information their are carrying.
- Such a schema has to be formally defined in RDF-OWL, using a specific tool suggested by iTelos (**Protège**).
- How to define a schema for the single dataset ?
 - The dataset structure is **self-explanatory**, thus reducing the modeling effort.
 - The dataset's **information needs to be interpreted**, thus a point of view is required for such an interpretation.
 - Which one ? **The Purpose**.

Inception - KGC

- The Knowledge Graph Construction activity takes in input:
 - (data layer resources) The datasets collected, cleaned and formatted.
 - (knowledge layer resources) The schemas produced (extracted) for each datasets.
- the objective of this activity is to create, for each pair composed by a dataset and its relative schema, a **single object representing a semi-formal information model**.
- Such an object is represented by a (dataset-specific) KG.
- To achieve this result a specific tool is offered by iTelos (Karma) used to map each dataset over its own schema, thus merging data and knowledge layer of a KG. ¹

¹This tool is exploited also in the last iTelos phase when the final KG is building

Contents

- 1 Phase structure
- 2 Inception - Input
- 3 Inception - Activities
- 4 Inception - In practice**

Inception - In practice

- The Inception phase is executed in practice by the following steps:
 - 1 (re)use and implementation (customization) of **data management libraries**, to execute the **scraping, cleaning** and **formatting** activities;
 - 2 datasets schema modeling using **Protège** ;
 - 3 KGC activity executed using the **Karma** tool².

Note: In the next lectures, demos will be presented for each tool mentioned.

²<https://usc-isi-i2.github.io/karma/>

Inception - In practice - Karma

■ Tool inputs:

- 1 Schema modeled for a single dataset;
- 2 dataset.

■ Tool outputs:

- JSON file: the datasets cleaned and formatted.
- RDF file: the KG-format object for the single dataset.
- TTL file: the instruction model to be used to reproduce the mapping activities
- dataset.

■ Git repository upload:

- initial dataset
- JSON file
- RDF file
- TTL file



KGE - Knowledge Graph Engineering



Fausto Giunchiglia



iTelos Methodology

Phase 1 - Inception