Defeating data scarcity by contextual reuse

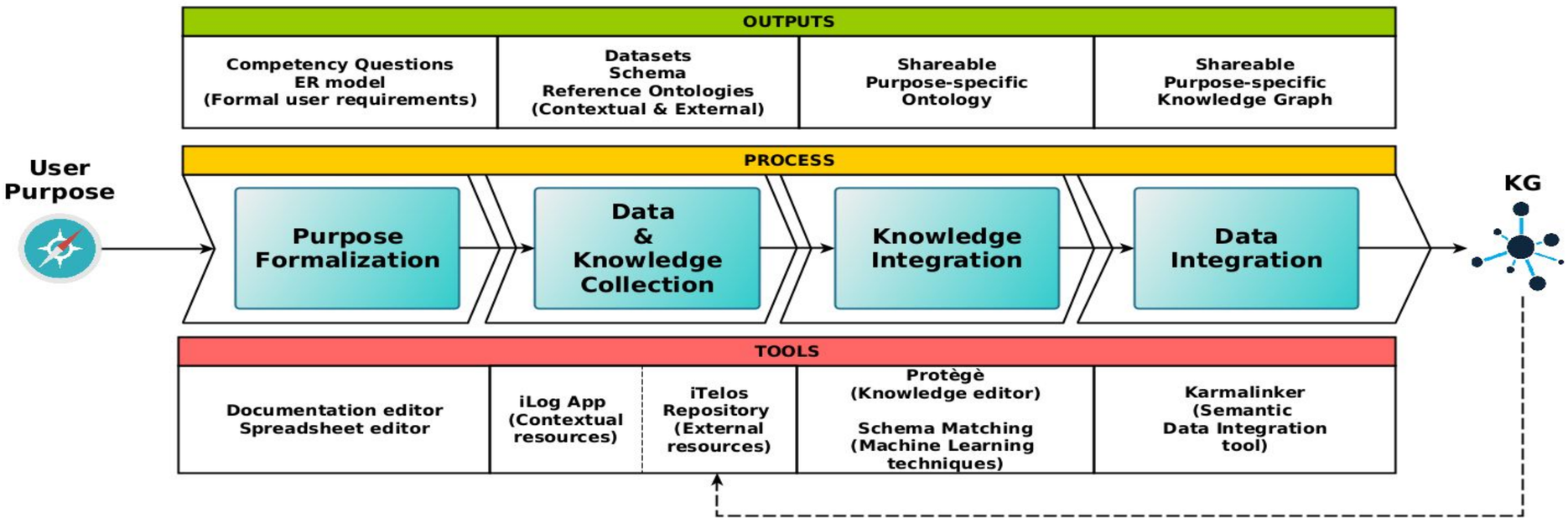
Simone Bocca (simone.bocca@unitn.it)
Matteo Busso (matteo.busso@unitn.it)

Department of Information Engineering and Computer Science
University of Trento

Motivation

The data and the information they brought are fundamental in every sphere of international cooperation, from research and social infrastructures to Industry 4.0. However, the information is only exploitable if both **data resources** and **knowledge resources** are recognized. These resources are often created according to the user requirements without considering further **reuse**. Without reuse, new projects are forced to produce new contextual, and hardly reusable, resources - a problem also underlying the well-known reproducibility crisis. In other words, resources are context-specific, i.e., exploitable in the same environment in which are produced. Users coming from different environments may not be able to exploit the same resources.
In this scenario, we propose a methodology, called iTelos, that facilitates data management-production actors in the production of context-specific resources - that is, based on their purpose - both via reuse of existing resources and via collection of new resources. However, iTelos goes beyond the mere production of resources: being based on a re-use centered process, its output will be highly shareable and reusable.

The iTelos Process



Purpose formalization: in this phase the user's purpose is formalized *into* a set of *Competency Questions* (CQ), i.e. user requirements. From CQ are selected those terms that better describe the Entity Types (ETypes) to be modeled into the final KG. The ETypes are *then used to build the purpose ER model* describing the information required to satisfy the user requirements.

Data & Knowledge Collection: has two objectives. (i) To retrieve all the existing resources suitable to satisfy the Purpose (here the iTelos repository plays a crucial role); (ii) to produce contextual resources exploiting the *data collection methodology* supported by the *iLog* mobile app, driven by the Purpose formalized in the previous phase.

Knowledge Alignment: this phase, aims to model the KG's unique ontology, called Entity Type Graph (ETG) using, e.g., the *Protègè* tool. The ETG is build by reusing existing knowledge resources (often well formed and standardized) matched using ML techniques. Parallelly, the external and contextual data resource are cleaned and formatted.

Data Integration: The ETG and the formatted data resources, are merged in a single data structure able to represent the resource involved. This phase exploits the Karmalinker semantic data integration tool, used to map the data resources on the ETG. The final output is stored in the iTelos repository (with metadata and documentation).

Two key ideas are implemented by the iTelos process.
1) **to enhance the reuse by producing reusable resources**: the process is supported by specific modelling methods and languages used to increase the shareability of the process outputs, as well as by a dedicated repository which plays both as a provider and final storage of the outcomes produced.
2) **to reduce the effort in purpose specific KG building:** the iTelos phases are structured to handle already existing data and knowledge resources with the aim of adapting them to the user's initial Purpose.

iLog (and the relative data collection methodology) is developed to collect data for many research purposes, both from the smartphone internal sensors and by sending context sensitive questions. It is adaptable to different purpose and easily downloadable on smartphones, which is an additional feature (the 83.96\% of the world's population owns a smartphone).

Use cases

iTelos - It has been partially adopted within InteropEHRate EU project [1], to produce interoperable healthcare resources in an EU cross-country environment. It is taught in the the Knowledge and Data Integration (KDI) master course [2], held in Trento (Italy) and Jilin (China). Students produced KGs (see Fig. 2) integrating external resources with exiting data, aiming at offering services for the university community. The student project's resources have been re-used in the every KDI edition. Students' feedback (see Fig. 1) leads improvements in the methodology.

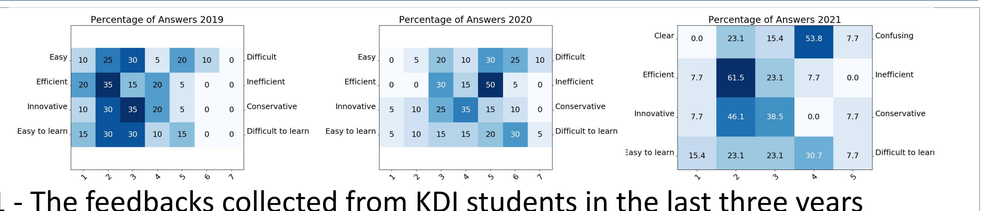


Figure 1 - The feedbacks collected from KDI students in the last three years

iTelos produce KGs that are concretely shaped as RDF files. RDF's structure allows the definition of the KG's ETypes (light blue circles in Fig. 2), the entities represented by those ETypes (yellow and red circles), as modeled in the KG's ETG, as well as the ETypes properties (arrows) used to associate data values (other circles) to each relative entity's attributes. It clearly appears how the RDF KG produced by iTelos, allows the exploitation of both the knowledge (ETG) and data (Entities and properties values) resources. iTelos KG can be presented, and used to support real services, through several tools and libraries able to handle RDF. The example reported in Fig. 2 shows the result of a SPARQL query execution using GraphDB tool.

Figure 2 – Graphical view of (Portion of) KG, Events project [3], KDI 2021.

iLog In WeNet EU project [4], iLog collected data about social habits of 600+ university students, for 1 month across three continents (see Fig. 3).
The iLog data collection methodology has been taught in the Studies on Human Behaviour (SHB) master course [5] in University of Trento (Italy), where students collected data about their daily routines, during the first COVID-19 pandemic year.

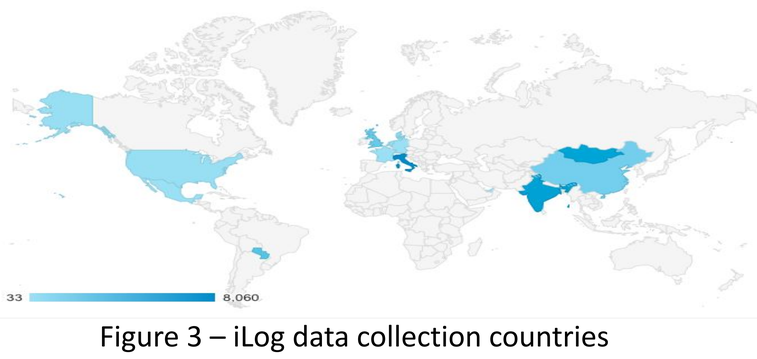


Figure 3 – iLog data collection countries

[1] https://www.interopehrate.eu/ [2]https://unitn-kdi-2021.github.io/unitn-kdi-2021-website/ [3]https://annafetz.github.io/ [4]https://www.internetofus.eu [5]https://unitn-shb-2020.github.io