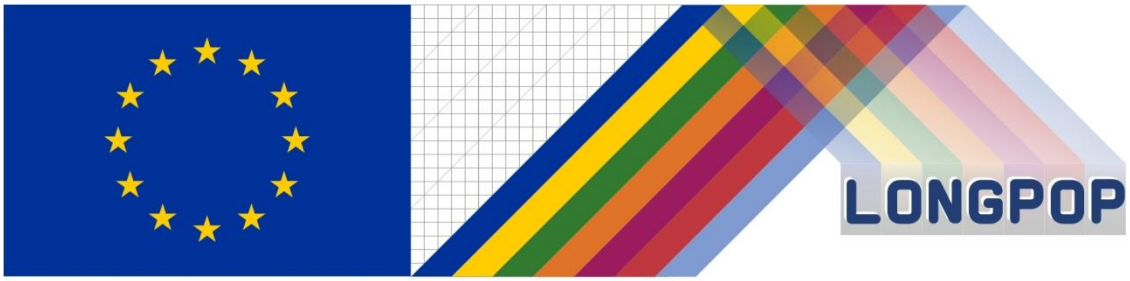


This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 676060

LONGPOP

Methodologies and Data mining techniques for the analysis of Big Data based on Longitudinal Population and Epidemiological Registers

Data Mining Extraction Software

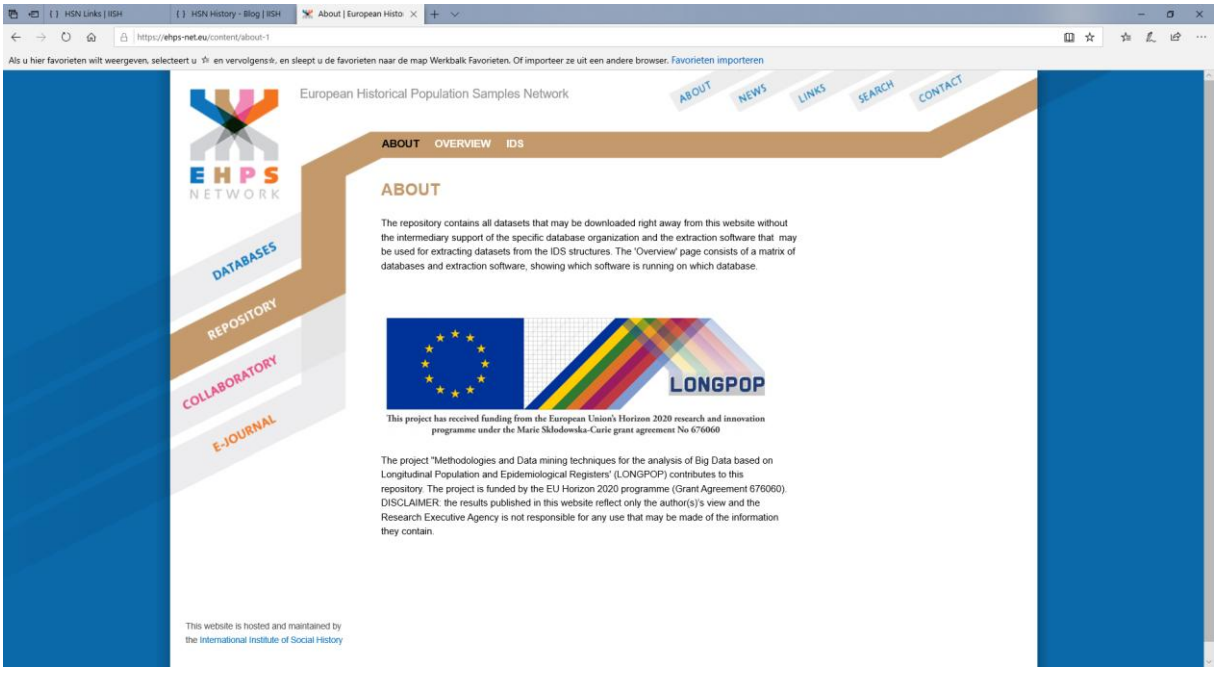
Deliverable n. 3.4

The Deliverable 3.4 of the MSCA ITN LONGPOP project entails the publication of extraction software based on the so-called Intermediate Data Structure (IDS; Alter, Mandemakers & Gutmann 2009; Alter & Mandemakers 2014). This publication structure, called *Repository*, was developed within a forerunner of the LONPOP project, the European Historical Population Samples Network (EHPS-Net) financed by the European Science Foundation Network.

The IDS is developed for and by scholars working with databases containing information on persons, families and households and forms an integrated and joint interface between all the databases that convert their data to IDS. The Intermediate Data Structure (IDS) was developed as a strategy aimed at simplifying the collecting, storing and sharing of historical demographic data (Alter & Mandemakers 2014; Alter, Mandemakers & Gutmann 2009). It is a common format that allows to manage and compare data from different databases, regardless of their original structure. IDS structures longitudinal databases and overcomes the problems of comparability which are inherent on large historical datasets on populations by providing a common dissemination format. In order to analyse the data contained in this intermediate structure, extraction software is developed to select the information from the IDS tables and to convert it into datasets ready for analysis. Since the requirements of each type of analysis vary, there will be many specialized extraction programs. This is an open, scalable, and extendable approach. An important issue is the documentation aspect which strengthens the development of extraction software in a systematic way.

So, one of the most substantive topic in the Work Package 3 was the further development of this extraction software. In the report WP 3.1 we have already given an overview of the software that has been developed within the LONGPOP project. This report D 3.4 confirms that the software has been effectively delivered through publishing by way of the Repository of EHPS net. This Repository can be reached by way of the following address:

https://ehps-net.eu/content/about-1



Alter, G., K. Mandemakers & M. Gutmann (2009**), Defining and Distributing Longitudinal Historical Data in a General Way Through an Intermediate Structure.** Historical Social Research 34, 3, 78-114.

Alter, G. & K. Mandemakers (2014). **The Intermediate Data Structure (IDS) for Longitudinal Historical Microdata, version 4.** Historical Life Course Studies 1, 1-26, published on line 26th of May 2014. PI: http://hdl.handle.net/10622/23526343-2014-0001?locatt=view:master