Navigating quality challenges in landscaping web data: New aspects and source stability

Magdalena Six, Alexander Kowarik[1]

Abstract

The experiences from the ongoing work in the ESSNet WIN show: No matter how well the process steps of data ingestion and data processing are done, the quality of the output very much depends on the quality of the source. As typical for Official Statistics, "quality of the source" is a multi-dimensional concept.
It can refer to several different aspects, e.g.: the stability of the access to the website; the availability of the most important information for the topic of interest; the trustworthiness of the website owner or the market share of the website, etc.
In this paper, we therefore focus on the process steps of finding out which web sources are available as input for a specific topic of interest and how to – if necessary – select the ones which will lead to the highest quality of the statistical product.
The paper is structured as follows: We first try to give a definition for the term Landscaping, and subdivide the process of Landscaping into three subprocesses "Catalogue", "Measure" and "Select". We analyse each subprocess and we show that the complexity of each subprocess is very use-case dependent. In the last chapters we focus on those cases where a selection model is needed. We first categorize groups of information upon which the selection model relies. We then present two selection models of varying complexity which were developed in WP2 and WP3.
The proposed actions might depend on national legalization. If the NSI is allowed to perform a certain action is not within the scope of this document.

[1] Statistick Austria, e-mail: Magdalena.Six@statistik.gv.at, Alexander.Kowarik@statistik.gv.at.