Item 5 – Web scraping and private sector data for official statistics

On guidelines for accessing Big Data at Eurostat

Fernando Reis

Eurostat A.5 – Methodology; Innovation in Official Statistics

Working Group on measuring e-commerce and the digital economy

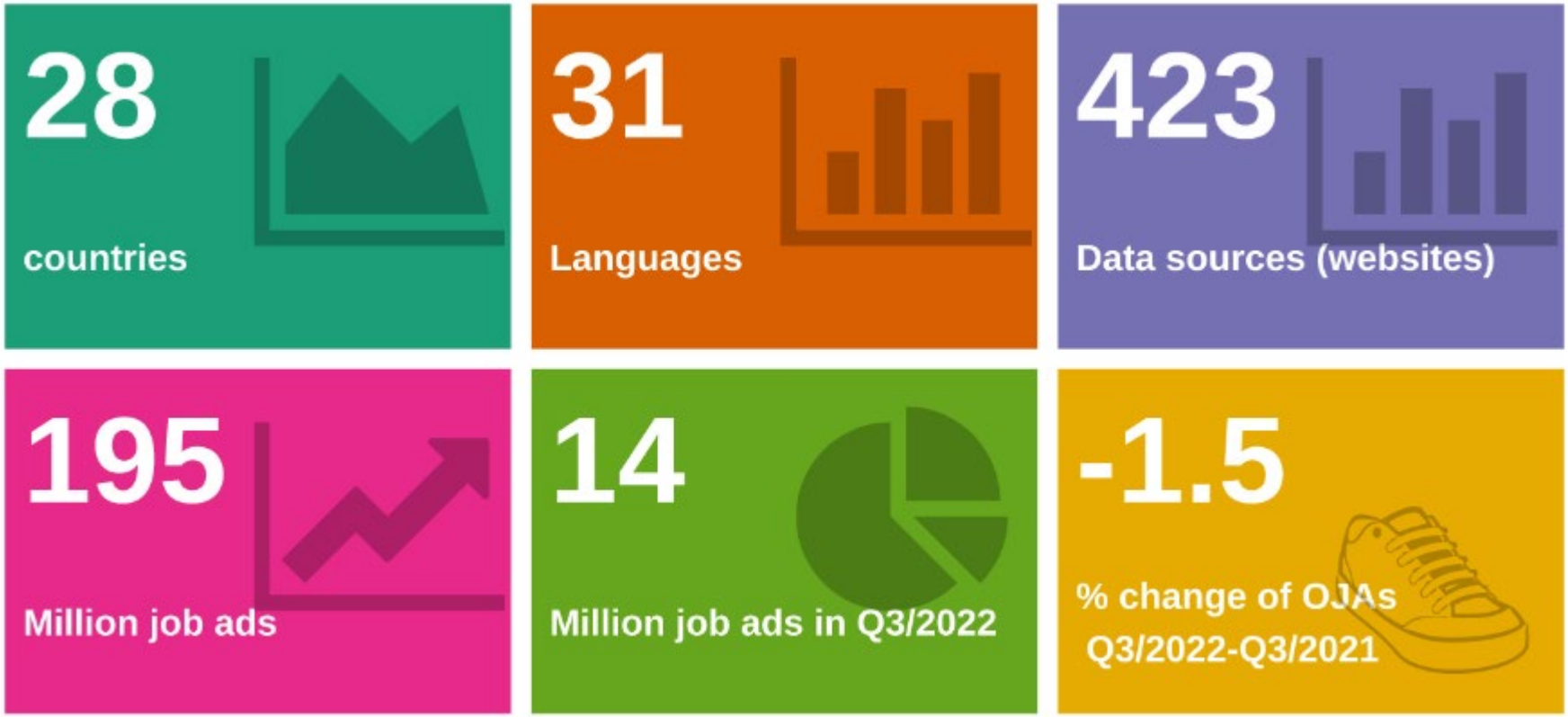
Third meeting 28-29 November 2022

Web Intelligence Hub (WIH) principles, rules and procedures of operation

Context

- **Web Content Retrieval Guidelines**
  - **Purpose:** harmonise practices in the ESS on the use of web content to produce statistics
  - Adopted on Feb 2022
  - https://ec.europa.eu/eurostat/cros/content/web-content-retrieval-guidelines-0_en
- **WIH principles, rules and procedures of operation**
  - Implement the web content retrieval guidelines in the WIH
  - Based on the experience with Online Job Advertisements

Online Job Advertisements



WIH rules and conditions

- WIH rules and conditions on web content retrieval
- WIH rules and conditions on web source agreements
- WIH rules and conditions on on access to content and data

- There will be more in the future, e.g.
  - **Dissemination of source code**

WIH rules and conditions on web content retrieval

WIH rules and conditions on web content retrieval

- Selection of Websites from which content is retrieved
- Identification of the WIH towards websites
- Crawling of the websites selected
- Minimisation of the impact on the web servers
- Treatment of sensitive content and data retrieved from websites
- Communication

WIH rules and conditions on web content retrieval

Selection of websites from which content is retrieved

- Landscaping
  - Inventory of websites
  - Content assessment
  - Selection based on quality and relevance
- Open commonly agreed methodology and objective selection criteria
- List of websites published
- WIH open to agreements

WIH rules and conditions on web content retrieval

Identification of the WIH towards websites

- User-agent string;
- List of IPs used by the WIH;
- *Web page with list of websites crawled;*

- Anonymous web crawling, of limited scope, performed for auditing purposes only

WIH rules and conditions on web content retrieval

Crawling of the websites selected

- Complies with robots exclusion protocol insofar as it is compatible with the legal mandate of the statistical offices to produce official statistics;
- If websites don't want to be crawled then they have to provide the data via other means;

WIH rules and conditions on web content retrieval

Communication

- General information about web content retrieval activities in web page (CROS portal);
- Advertise:
  - these rules and conditions;
  - the list of use cases run under the WIH;
  - the list of websites selected in each use case;
  - the landscaping methodology used in the selection of websites;
  - the list of statistics produced with the content retrieved;
- Informs website owners individually when content retrieved has a significant impact on the web server
- Contribute to create communities of data sources

WIH rules and conditions of source agreements

WIH rules and conditions on web source agreements

- Types of agreements
- Criteria for establishing an agreement with the WIH
- Conditions that can be defined in a content retrieval agreement
- Communication

WIH rules and conditions on web source agreements

Criteria for establishing an agreement with the WIH

- Selection of website by a landscaping;
- All owners of website selected can have an agreement;
- The WIH will communicate and engage with the website owners according to the importance of the website for the data collection, based on a score defined in the landscaping;
- Top most important websites may be contacted for a cooperation agreement;

WIH rules and conditions on web source agreements - Web content retrieval conditions that can be defined in the of agreement

The WIH establishes agreements with website owners for the benefit of both.

When establishing agreements, the WIH seeks:

- An uninterrupted retrieval of content from the website to the WIH;
- More efficient channels of web content retrieval such as the use of APIs;
- Possibility of provision of additional data not available in the website. These data will be considered statistical confidential and will be available only to producers of official statistics under legal provisions guaranteed in statistical legislation. (see WIH Data Access Policy);

WIH rules and conditions of source agreements - Web content retrieval conditions that can be defined in the of agreement

With the agreements, the WIH will offer to the website owners:

- A reduction of the impact of the content retrieval in the web servers, by agreeing specific conditions;
- Web source portal in the platform with dashboards to monitor the web content retrieval;
- Publicity / Visibility - registry of job portals (more or less visible);
- Analytical services, e.g. API for classification;

WIH rules and conditions of source agreements – Communication

- Advertise on the Web
  - these rules and conditions;
  - list of websites with agreement - specific conditions of agreements not undisclosed
  - list of websites selected in each use case, inviting owners of the websites to contact the WIH for the establishment of an agreement, if they wish so
- Direct contact of the WIH with most important websites