ESSnet Trusted Smart Statistics – Web Intelligence Network

Grant Agreement Number: 101035829 — 2020-PL-SmartStat

WP2: OJA and OBEC Software

Deliverable 2.5 List of requirements defined by the WISER group with the result of its implementation

Version 1.0, 2025-03-31

Prepared by:

WP leader: Jacek Maślankowski (Statistics Poland, Poland, j.maslankowski@stat.gov.pl)

Contributors:

Massimo De Cubelis, Alessandra Righi

Web Intelligence Network

Funded by the European Union

This document was funded by the European Union.

Contents

1. Introduction

One of the goals of the Web Intelligence Network (WIN) was to establish a set of requirements to enhance the functionality of the Web Intelligence Hub (WIH), as well as the Online-Based Enterprise Characteristics (OBEC) and Online Job Advertisements (OJA) use cases. Further details on these use cases and the WIH can be found in the references of this deliverable (WIN WP2, 2022; WIN WP2, 2023; WIN WP2, 2024; WIN WP2, 2025).

The requirements were contributed by both WIN members and the WISER group—a panel of potential users of the WIH and its associated use cases, including the Data Acquisition Platform and OJA Datalab. These requirements were gathered during three workshops held on 7 May, 14 May, and 4 June 2024, conducted via the Webex videoconferencing tool. Additionally, supplementary requirements were collected through a questionnaire, which is included in the Annex of this document.

2. Requirements

2.1. Web Intelligence Hub

The table below shows the requirements related to the Web Intelligence Hub (WIH) defined by the WISER group and responses from the Web Intelligence Network community related to them.

Table 1. Requirements related to the Web Intelligence Hub

Requirement	Description	Implementation status
R_WIH.1	The WIH is too large and complicated for small use cases and could be simplified	There is a confusion what is the WIH and its components. On the new CROS portal there is a dedicated section for the WIH to provide an overview and some additional details to some specific subjects. In addition, some “hello world” examples for the data acquisition service (DAS) will be added to the documentation.

R_WIH.2	The WIH could check whether mobile / desktop versions of the websites are available.	The website can be rendered to specific web browser / version with Selenium, so it is possible to extract this feature by web scraping to different web browsers and checking the results.
R_WIH.3	The WIH could register loading time of the website (response time).	The response time is not a good measure because it very much depends where are the server located, what is the size of the computer used for crawling and dynamically or static we retrieve the given page.
R_WIH.4	The WIH could check e-commerce attribute of the website.	It is possible to extract e-commerce attributes from the downloaded website HTML in the Datalab, which allows to have different perception of different perception of e-commerce definition used in different use cases / surveys.
R_WIH.5	The WIH should have one available in one place documentation from A-Z.	Currently the documentation is spread in 3 places (WIH wiki, DAS platform docs and datalab docs). In the future these documentation will be merged.
R_WIH.6	The WIH should have regular updates of documentation (some scripts are not working).	There is a WIH newsletter, where the latest updates are mentioned. The newsletter also includes updates to the DAS and Datalab

The WISER group defined several requirements for the Web Intelligence Hub (WIH), and the WIN community responded with clarifications and planned actions:

- Simplification (R_WIH.1) – There is confusion about WIH components. A dedicated section will be added to the new CROS portal, along with "hello world" examples for the Data Acquisition Service (DAS).
- Checking Mobile/Desktop Versions (R_WIH.2) – The WIH can determine website versions using Selenium for web scraping.
- Measuring Loading Time (R_WIH.3) – Not recommended due to dependency on server location, hardware, and page retrieval method.
- Checking E-commerce Attribute (R_WIH.4) – Feasible only in Datalab, as e-commerce definitions vary across use cases.
- Comprehensive Documentation (R_WIH.5) – Currently spread across three sources (WIH wiki, DAS docs, Datalab docs). Plans to merge them into one unified source.
- Regular Documentation Updates (R_WIH.6) – Newsletter updates will address documentation changes, with a major update planned for Q1 2025.

2.2. OJA Use Case

The table below shows the requirements related to the Online Job Advertisements (OJA) Use Case defined by the WISER group and responses from the Web Intelligence Network community related to them.

Table 2. Requirements related to the OJA Use Case

Requirement	Description	Implementation status
R_OJA.1	The OJA datalab should deliver the current list of data sources used to generate statistics.	The list of sources are not available due to confidentiality. Only anonymized statistics can be retrieved for the sources in the Datalab.
R_OJA.2	The OJA datalab should have more detailed metadata (e.g. whether ML method was used to identify occupation).	Currently the general overview is included in the documentation.

R_OJA.3	The OJA datalab should have documentation in one place with regular updates of the code in R, Python, SQL to get the data.	Eurostat updates the documentation after major changes. The OJA datalab documentation will be merged with the the documentation of other components.
R_OJA.4	The OJA datalab should include quality indicators (e.g. accuracy of ML algorithm).	Some confidence measures are calculated when ML algorithms are used. This information can be retrieved on request.
R_OJA.5	The OJA datalab should have an attribute of public / private sector breakdown.	It is not possible due to insufficient data.

The summary of the table shows that there are some requirements that cannot be implemented due to the limitations of the OJA data sources:

- R_OJA.1: The OJA datalab cannot provide an on-demand list of data sources to external users, but Eurostat can deliver it upon request.
- R_OJA.2: More detailed metadata (e.g., ML methods for occupation identification) is needed; currently, only a general overview is available in the documentation.
- R_OJA.3: Eurostat is working on consolidating documentation and providing regular updates for code (R, Python, SQL) to access data.
- R_OJA.4: Quality indicators (e.g., ML algorithm accuracy) are currently unavailable in the datalab.
- R_OJA.5: A public/private sector breakdown cannot be included due to insufficient data.

2.3. OBEC Use Case

The table below shows the requirements related to the Online Based Enterprise Characteristics (OBEC) Use Case defined by the WISER group and responses from the Web Intelligence Network community related to them.

Table 3. Requirements related to the OBEC Use Case

Requirement	Description	Implementation status
R_OBEC.1	The OBEC indicators should consider loading time / mobile version of the application.	It is not possible as the server who is scraping can be located in different places, affecting the loading time.
R_OBEC.2	The OBEC indicators should also focus on technical quality of web pages.	It is possible to get this information by parsing the website in Datalab. This issue was described in the deliverable 2.4, e.g., how to detect and get data extracted directly from the HTML markups, like lang technical parameter.
R_OBEC.3	The OBEC documentation should clearly explain what e-commerce is.	Implemented. Deliverable 2.4 includes the final definition of SMP and E-commerce.
R_OBEC.4	The OBEC indicators can also include multilanguage support of the webpage.	Implemented. Deliverable 2.4 explains the methodology and script has been delivered on internal project Gitlab.

In contrast to the OJA, OBEC requirements were more easy to be adapted, except one of them (R_OBEC.1):

- **R_OBEC.1**: Loading time/mobile version assessment for OBEC indicators is not feasible due to varying server locations affecting measurements.
- **R_OBEC.2**: Technical quality analysis of web pages is possible via parsing in Datalab; the code was explained in the Deliverable 2.4.
- **R_OBEC.3**: The documentation now clearly defines e-commerce (included in Deliverable 2.4).
- **R_OBEC.4**: Multilingual webpage support is addressed—methodology is in Deliverable 2.4, and the script has been uploaded to the internal Gitlab.

3. Summary

Requirements defined by the WISER group show that there is a interest of external users in using the Web Intelligence Hub and use cases developed by WIN. However, some requirements cannot be satisfied due to lots of limitations related to the data sources used in use cases, as well as the platform itself. It is highly recommended that future work related to WIH, OJA and OBEC should be related to the requirements defined in this documents.

According to the WISER group, the benefits on the use of the platform is its scalability and the possibility to scrape the data with the newest libraries, e.g. selenium with possibilities to render to different devices. WISER users were also very satisfied with the scope of the OJA database. Users from Spain mentioned that this is one of the data sources they can use to extract the data at regional level. The most useful indicators were related to OJA, i.e. results by skills and occupations with regional disaggregation. From the OBEC perspective, users were mostly interested in e-commerce related indicators as well as multilanguage support of the website.

List of tables

References

Web Intelligence Network: WP2 Deliverable 2.1. First Interim Progress Report, 2022.

Web Intelligence Network: WP2 Deliverable 2.2. Second Interim Progress Report, 2023.

Web Intelligence Network: WP2 Deliverable 2.3. Third Interim Progress Report, 2024.

Web Intelligence Network: WP2 Deliverable 2.4. Final Progress Report, 2024.

Annex 1. QUESTIONNAIRE 1

FINAL VERSION OF THE WIH PLATFORM

Meeting: Presentation of the final version of the platform + training

J. Maślankowski, M. Meszaros – 45 min.

Date: 13th May 2024

QUESTIONS

Section 1 – Most useful functionalities

What functionality could be most useful for your future use case?

……

……

Would it be applicable as it is or could it be improved? Yes __ No __

If yes, how?

……

……

Are there other functionalities you find useful that could be improved? Yes __ No __

If yes, which ones?

……

How would you improve them?

……

Section 2 – New functionalities to propose

Are there functionalities that could be useful but are not included in the current version of the software?. Yes __ No __

If yes, which ones (please indicate at least two)?

……

……

Could you describe the input data, the procedures, and the desired output?

……

……

Section 3 - Usability

On a scale from 1 to 10, how user-friendly do you find the software for the following features to be (1 being not at all user-friendly - 10 being very user-friendly)?

	1	2	3	4	5	6	7	8	9	10
data acquisition										
filtering function										
performance										

How would you improve the features you find less user-friendly?

……

……

Section 4 – Benefits

What are the benefits from your point of view in using the WIH platform?

……

……

Annex 2. QUESTIONNAIRE 2

Meeting: OJA Datalab hands-on training on preparing tables with experimental data

J. Maślankowski - 90 min.

Date: 7th May 2024

QUESTIONS

Section 1 – Functionalities and Procedures

On a scale from 1 to 10, how user-friendly do you find the procedures for the following functionalities to be (1 being not at all user-friendly - 10 being very user-friendly)?

	1	2	3	4	5	6	7	8	9	10
data acquisition										
filtering function										
performance										

What features do you find to be not user-friendly and how would you improve them?

……

……

Section 2 – Experimental Outputs

What do you consider the most useful indicators about quarterly changes that could be generated with OJA information? (Please, rank the proposed indicators from 1 (most useful) to 5 (less useful))

Indicator	**rank**
Occupation by countries and quarters	
Skills by countries and quarters	
Fluctuation of OJAs by countries and quarters	
New OJAs by countries and quarters	
Most demanded occupations / skills by countries and quarters	

At which level of breakdown do you consider more useful the set of tables/indicators proposed to measure quarterly changes**?**

- regional or national
- 5 occupations
- including time series, metadata, quality aspects (skills specific for these 5 occupations), accuracy indicator
- at which level of NACE rev.2 classification?

Are you or your organization interested in other indicators? Yes __ No __

If yes, which ones?

……

……

Detailed questions related to OJA data:

1. Do you find it useful to add information on precision/recall/F1-score/confident etc. from ML algorithm on each occupations?

Yes / No

Please explain your decision:

……

……

2. What additional variables you want to add to OJA data, if any?

……

……

3. Do you find it useful to create dashboards in Datalab with predefined tables, with the most reliable occupations/skills/etc?

Yes / No

Please explain your decision:

……

……

If yes, please tell us what variables / breakdowns should be included?

……

……

4. Do you want us to share the scripts with predefined tables to get the most recent data from OJA DataLab (e.g., occupations by countries, time series of skills by countries, the most demanded skills etc.)? If yes, tell us which one are the most important (e.g. time series, what breakdowns) for you and what language you prefer – R or Python?

……

……

5. Can you define more requirements / changes / additional products you would like to get from OJA Datalab?

……

……

6. How likely is it for you to use this dataset in the future?

……

……

Annex 3. QUESTIONNAIRE 3

Meeting: OBEC training on the process of data collection/ processing/analyzing on the WIP

J. Maślankowski - 90 min.

Date: 4th June 2024

QUESTIONS

Section 1 – Functionalities and Procedures

On a scale from 1 to 10, how user-friendly do you find the procedures for the following functionalities to be (1 being not at all user-friendly - 10 being very user-friendly)?

	1	2	3	4	5	6	7	8	9	10
data acquisition										
filtering function										
performance										

What features do you find to be not user-friendly and how would you improve them?

……

Section 2 – Experimental Outputs

What do you consider the most useful indicators that could be generated with information shown on the enterprises' sites? (Please, rank the proposed indicators from 1 (most useful) to 5 (less useful)

	rank
Social media presence	
E-commerce	
Chatbot	
Multilanguage support	
Extracting contact information	

Are you or your organization interested in other indicators? Yes __ No __

If yes, which ones?

……

……