International Labour Organization

▶ ILO brief

▶ Research Brief

December 2024

A new chapter for the ILO's textual assets

Applying Generative AI to Labour Force Survey questionnaires[1]

By Pawel Gmyrek and Poleth Vega Ruales

Key points

- This note documents the use of Generative AI to create a database of survey questions at the ILO, based on ILO micro data collection.
- We demonstrate how AI vision capabilities can be effectively combined with human verification to ensure accurate data extraction from complex survey forms.
- We use several Generative AI tools to facilitate the digitization, storage, and semantic search through vector-based embeddings of Labour Force Survey question-answer pairs, improving accessibility and efficiency for ILO staff and external researchers. By deploying Generative AI models, we automate the search function and generation of research summary notes, streamlining the process for users with varying levels of technical expertise.
- We provide an interactive search app to ILO-authorized users. In addition, we provide a GitHub repository with step-by-step instructions and a full Python code.
- We suggest that there is major potential for the application of this process to other textual resource at the ILO. We argue for the need to develop strategies to turn such resources into valuable digital assets for the future that can be leveraged for research and policy development, and which can provide long-term value as the abilities of AI technologies evolve.

[1] This project was conceived and funded by the Effective Labour Institutions (ELI) unit of the ILO Research Department. We thank Janine Berg for her support and very helpful comments. We thank Steve Kapsos for his review of this brief and for organizing a series of constructive discussions with staff members of ILO Department of Statistics. We also thank Judy Rafferty and Pierre-Alain Proust for their support on the layout and publishing process of this note.

Introduction

For over 100 years, the International Labour Organization (ILO) has been gathering a vast collection of written documents related to the world of work. While many have been digitized or were initially produced using digital technologies, they are largely unstructured and stored in disparate file formats (.pdf, .docx, .xlsx). This lack of structure makes it cumbersome to access for research or other purposes. In this brief, we demonstrate how the new generation of publicly available AI technology can be applied to transform such documents into a "textual asset" that can facilitate the work, including by increasing both speed and accuracy, of ILO staff. By using the example of labour force survey forms from the ILO micro data repository, we demonstrate a chain of techniques applied to transforming raw textual documents into a rapidly searchable database and an accompanying tool to summarize findings.

Our efforts result in the creation of a new database of 30,691 questions pertaining to labour force surveys (LFS) from 132 countries[2], which is accessible to ILO staff and authorized external researchers. While this is a valuable product on its own, we rely on this technical note to explain key concepts in detail and document our steps. Ultimately, we aim to encourage more ILO colleagues to experiment with Generative AI (GenAI) tools to make existing information more accessible and useful for researchers and policymakers.

We rely on the GenAI models available from OpenAI libraries for the purpose of the work demonstrated in this brief. This selection is solely due to the authors' greater familiarity with OpenAI libraries and the current capabilities of its individual models. The authors of this text have no affiliation to OpenAI and no vested interest in that regard. In practice, the same work could be performed with a range of other commercial GenAI models available through an Application Programming Interface (API). It is also possible that some of the open-source models could achieve a similar result, especially if the operators are able to fine-tune them to their needs.

Digitization of Labour Force Survey questionnaires: A case study

The ILO Department of Statistics collects household survey datasets, primarily labour force surveys (LFS), from over 160 Member States. This extensive repository of survey questionnaires is pivotal for supporting global research, policymaking, and the monitoring of labour market trends. However, traditional methods of managing and analysing these questionnaires pose significant challenges due to their time-consuming and cumbersome nature.

For example, consider a researcher wanting to find identify survey questionnaires that ask the respondents about workplace harassment. Under the current system, this task involves manually navigating through individual folders of the micro data repository, opening each questionnaire stored on the ILO's internal drive and conducting separate keyword searches. Survey questionnaires are mostly stored as .pdf, .docx, .xlsx files, and can sometimes be composed of separate file elements. Searching for questions referring to the notion of "workplace harassment" typically requires building a lexicon of associated terms, which are then entered into a keyword search performed on each document separately. While parts of this work can be automated with Natural Language Processing (NLP) techniques, it remains a cumbersome iterative process. Moreover, automating with NLP tools requires a solid understanding of programming languages like R or Python, which can be a barrier for researchers without a quantitative background.[3]

In this context, the integration of digitization and semantic search optimization with the use of GenAI tools emerges as a transformative solution to enhance the efficiency, accuracy, and accessibility of these data resources to a wider range of users with different digital skills and research interests. In the following sections we describe in detail each step in this process.

[2] The following countries from the ILO microdata have not been processed yet and could be added to the database at a later stage: Australia, Belarus, Brazil, Botswana, Congo, Comoros, Estonia, Nicaragua, Niue, Nauru, Palau, Papua New Guinea, Russian Federation, Sudan, Solomon Islands, El Salvador, South Sudan, Sao Tome and Principe, Suriname, Chad, Tokelau, Turkmenistan, Trinidad and Tobago, Tuvalu, Ukraine, Uruguay, Uzbekistan, Venezuela (Bolivarian Republic of), Vanuatu, Wallis and Futuna, South Africa.

[3] ILO STATISTICS offer harmonized micro data based on relevant questions from individual questionnaires. However, harmonized data does not provide a direct link to the original phrasing of the question in the country-level questionnaire, which is often of interest to researchers focusing on exploring specific theme, like "workplace harassment" in our example. In addition, harmonized datasets focus only on the questions that were processed by ILO STATISTICS and not on all the questions available in the original questionnaire.

Step 1: AI vision: Structuring and verifying documents into Q&A data frames

We rely on the image recognition abilities of GPT-4o and process each survey form individually. The first step is to convert a .pdf or another file format that holds the original questionnaire into a collection of images, each representing a single page of the original document. These images are stored in a temporary folder on the drive in .jpeg format, with names corresponding to page numbers in the original file. The next step involves encoding these image files into a base64 format, as required by GPT-4o for image input.

Next, we design a loop that sequentially processes each .jpeg image in the folder, sending them to the API of GPT-4o with the following instructions[4]:

```
"text": "This is a page from a survey questionnaire."
        "Retype all questions and their answer options into separate lines."
        "Mark each question with a Q and each answer option with an A."
        "Keep question numbers at the start."
        "Make sure to not invent content - if you cannot recognize the image, just state it and write 'error'."
```

The loop is designed to capture all processed pages into one .docx file. The specific formatting instruction – to mark each question with a "Q" and each answer with an "A" – results in a neat output that can be easily processed further in an automated way. Our pre-tests also show that providing specific and detailed instructions to the model on how to handle problematic images greatly reduces the incidence of hallucinations.[5] As the final step, we proceed with human verification of each .docx file, which looks as in the below illustration:

```
Q5: Has any child under the age of 5 years died in this household in the past 12 months?
A5:
01: Yes
02: No

Q6: In the past 30 days did you or any household member eat fewer meals in a day because there was not enough food because of a lack of resources?
A6:
01: Yes
02: No
```

Human verification consists of a meticulous comparison of the AI-generated output to the original survey file. In many cases, the model demonstrates an astonishing ability of recognizing text and making logical connections, for example, by matching questions presented in a complex table to answer options that were provided in a footnote at the bottom of a survey questionnaire. In other cases, correcting the output can require substantive human effort, and such corrections can vary from misnumbered question references to truncated or omitted text sections.

While human verification is a tedious and time-consuming process, it is crucial for producing high-quality data through human-AI collaboration. The use of AI vision as a starting point makes certain tasks feasible that would otherwise be too costly and labour-intensive to undertake. For example, in the first two months of our project, we processed 158 survey documents comprising a total of 30,691 individual survey questions. We estimate that approximately 60 per cent of the text was correctly processed by the AI on the first attempt. Verifying each question individually was a significant effort, but it was far more feasible than retyping all survey questionnaires manually to create structured files. Importantly, this verification effort is a one-time task: once the content is verified, it can be reused in multiple ways, as demonstrated in the following sections.

[4] Full Python code available upon request.

[5] In this context, "hallucinations" refer to the AI generating information that is incorrect or not based on the input data.

After verification, the .docx files are converted into .xlsx structure by a Python-based text processing loop. Specifically, this loop uses the question ("Q") and answer ("A") markers to separate the content into columns, with the country variable automatically added to each file and file name. This results in a collection of .xlsx files, each representing a survey questionnaire and structured as per the below illustration:

▶ **Table 1. Structure of question-answer pairs after AI processing and human verification**

country	question	answers
AUT	**Q1. Is that self-disclosure or third-party disclosure?**	A1. Person provides the information themselves
		A2. Fremdauskunft (Eng. "information provided by a third party")
AUT	**Q2. Did you live at the same address as now a year ago (on .....)? (Date of the Sunday of the reference week one year ago)**	A1. Yes
		A2. No
AUT	**Q3. Did you live in the same federal state at that time (.....)? (date of the Sunday of the reference week a year ago)**	A1. Yes
		A2. No, in another state – Continue with B3a
		A3. No, in another state – Continue with B3b

Finally, such files can be automatically pooled into one database, with additional variables added to specify date, survey version and similar relevant parameters, based on the original mapping of the selection in the ILO micro data repository. Importantly, in this final step, we add variables that allow for a quick identification of the original survey file on the ILO drive – such as the path and file name – which become part of the metadata, as demonstrated in the following section.

Step 2: Building a database for rapid semantic searches

In the second step of the project, we rely on a different characteristic of GenAI tools, which is to convert text into its numeric representation, known as "embeddings". This feature enables rapid semantic searches. While there are many technicalities to this concept, in straightforward terms, an embedding is an array of numbers, of which each number represents one dimension of a highly dimensional vector, and where the entire vector represents a section of text. In the case of OpenAI embeddings, such vectors have a standard length of either 1536 or 3072 dimensions; other LLMs rely of other idiosyncratic lengths.

Embeddings can be requested directly from OpenAI and stored in a third-party vectoral database, which offers significantly more rapid content search possibilities than a database with only text context. In addition, thematic searches can be performed based on semantic proximity of the searched concept to the content of the database, as opposed to more traditional methods of searching text, for example based on keywords and lexicons of key terms, as will be shown in the example below.

In mathematical terms, comparing the semantic proximity of two pieces of text that are stored in a vectoral (embedding) form can be performed through one of the basic methods of calculating vector similarity. We use the most popular method of calculating cosine similarity. Such similarity scores range from -1 to 1, with 1 representing closest proximity between two pieces of text in semantic terms. Illustrating semantic search with the previously discussed concept of "workplace harassment", the embedding corresponding to this term takes the following form:

"workplace harassment" = ([-0.01167822, -0.0396051 , -0.00396611, ..., -0.01767341, [*another 1530 numerical elements*], 0.01623265, 0.01462379])

In trying to find the three most closely related question-answer combinations in our LFS database, we calculate the cosine similarity of each embedding in the database to the one representing the search term and, subsequently, sort the search results descending. The top three searches are most closely related to "workplace harassment" and listed in the table below:

▶ **Table 2. Result of a semantic seach for "workplace harassment", with similarity scores**

year	source	question	answers	country name	similarities
2023	LFS	Q27.c Has (NAME) ever verbally or physically abused at his/her workplace?	A1. Yes\nA2. No	Indonesia	0.452894
2019	LFS	Q13_14. In the last 12 months, have you ever been or felt subjected to any of the following at work or by customers? Please choose the MAIN one.	A01. Constantly shouted at A02. Repeatedly insulted A03. Beaten/physically hurt A04. Touched or done things to you that you did not want A05. Gender discrimination A06. Ethnic or regional discrimination A07. Discriminated for being of foreign origin A08. Sexual harassment A09. Others (Specify) A10. NONE	Cook Islands	0.440822
2022	LFS	QOSH_07. During the last 12 months, have you been subjected to the following at work?	A1. Constantly shouted at/ repeatedly insulted A2. Beaten /physically hurt A3. Sexually abused (touched) A4. None of the above A5. Other (specify). _____	Bangladesh	0.437829
2023	LFS	Q14. OSH_ESF In the last 12 months have you ever been subject to the following at work?	A. a. Constantly shouted at A. b. Repeatedly insulted A. c. Beaten/physically hurt A. d. Sexually abused (touched or done things that you don't want) A. e. NONE A. x. OTHER SPECIFY _____	Eswatini	0.434687
2021	LFS	Q H10 In the last 12 months, have you ever been subjected to the following at your workplace? (Read each of the following options and mark "YES" or "NO" for all options)	A A. Constantly shouted at A 1=Yes A 2=No A B. Repeatedly insulted A 1=Yes\nA 2=No A C. Beaten /physically hurt A 1=Yes A 2=No\nA D. Sexually abused (touched, said or did things to you that you did not want) A 1=Yes\nA 2=No\nA E. Financial Abuse (Not paid for services offered)\nA 1=Yes A 2=No A Z. Other (Specify)..... A 1=Yes A 2=No	Uganda	0.42658

This process can be automated further, by relying on basic properties of an online vectoral database. First, we convert the entire structure of our survey questions database into a JSON format, with descriptive columns used as metadata, and each pair of question-answer options treated as the main text content. Hence, an individual record takes the following form:

{'country': 'AFG',
'region': 'Asia and the Pacific',
'income': 'Low Income',
'year': '2021',
'source': 'LFS',
'question_index': 22,
'question': 'Q1.18 Interview finish time',
'answers': 'A Hour\nA Minute\nA AM/PM',
'questionnaire_filename': 'AFG_LFS_2021_Questionnaire.xlsx',
'path':
'J:\\DPAU\\MICRO\\AFG\\LFS\\2021\\ORI\\Questionnaire\\','full_path':'J:\\DPAU\\MICRO\\AFG\\LFS\\2021\\ORI\\
Questionnaire\\AFG_LFS_2021_Questionnaire.xlsx',
'filename_processed': 'AFG.xlsx',
'all_text': 'Q1.18 Interview finish time A Hour\nA Minute\nA AM/PM'},

Note that, in addition to the JSON conversion, we created a new category, "all_text", which combines both the question and the corresponding answer options into a single text block. The advantage of this approach is that we can use this combined text as the basis for embeddings, enabling simultaneous search for semantic proximity in both the question and the answer block. This is crucial because the concept of interest may sometimes only appear in the answer options rather than the question itself.

Indeed, this is the case for Bangladesh LFS in table 2, where the relevant concept for workplace harassment is present in the answer block, yet none of the terms use the word "harassment". This highlights the superiority of semantic search methods over traditional keyword-based searches, which would require a comprehensive lexicon to capture related terms such as "shouted at" or "beaten." Using embeddings eliminates the need for such keyword lists, as concepts are matched based on vector proximity, which captures semantic similarity and works across languages. In this case, terms like "shouting" and "beating" are semantically close to "harassment" and "violence," and are therefore picked up effortlessly by the search engine.

After preparing the text, we enter the elements of this new JSON file into Weaviate[6], which is a cloud-based rapid vectoral database. Weaviate automates the function of converting search terms into embeddings and calculating cosine similarity scores to identify the closest matches. Semantic searches in Weaviate return content in the form of a JSON dictionary, which can be treated either as the final result or used for further processing with Generative AI (GenAI) tools, as demonstrated in the following section.

Step 3: Developing a GenAI summarizer for automated insights

Since Weaviate searches return text content, we can leverage the abilities of GenAI models to process such text further. In our case, we use GenAI to automate the development of a final note that the user can download in a fully formatted Word file. This is done by passing the result of the search to the API of GPT-4o, alongside additional instructions for the model – such as the searched term – and a request to provide a succinct summary. The detailed prompt is constructed as follows:

[6] Weaviate is only one of the possible technical solutions for a vectoral database. One of its advantages is that it holds the original text alongside the embeddings, which enables easy verification of the retrieved content. ILO/ITC Turin developed a range of services and systems, related to their quickly expanding expertise of building custom-made chatbots, for which such vectoral databases as a key element of the design.

```
response = client.chat.completions.create(
  model='gpt-4o',
  messages=[
    {"role": "system",
    "content": "You will by a helpful research assistant to a labour lawyer."},
    {"role": "user",
    "content": "Look at this text:" + str(search_result) +
    "It is the result of my semantic search of vectoral database with questions from different national surveys" +
    "I searched for the concept" + concept +
    "Based on this search result, write a well-structured analytical summary note about how the concept of "
      + concept +
     "is reflected in national surveys."
    "Support your statements with references to the content of country surveys from the search."}
    ])
```

In order to make the process more user-friendly, as the final step we develop an interface based on a simple R Shiny format. This means that the technicalities described in Step 1-3 are hidden in the R and Python scripts operating behind the app, while the user is able to establish search parameters (such as the number of most relevant questions or the region), conduct the search and produce the final note without any knowledge of programming. The app is available for ILO users under the following link[7]: https://pgmyrek.shinyapps.io/LFS_Questions_Database/

AI-powered analysis of processed LFS questions

Our database stores 30,691 processed question-answer pairs from the LFS surveys alongside metadata and associated embeddings, enabling text analysis using a combination of traditional NLP methods and GenAI-powered techniques. Figure 1 provides an overview of the country coverage.

The 132 countries currently covered by the database are organized into ILO regions (Africa, the Americas, the Arab States, Asia and the Pacific, and Europe and Central Asia) and World Bank income groups (high income, upper-middle income, lower-middle income, and low income). Figure 1 reveals significant variation in the number of question-answer pairs across countries and regions, with China having the largest number of questions (1,396) and Japan the lowest (27).

Since embeddings represent semantic concepts, we can leverage their mathematical properties to explore thematic groupings within the LFS forms. This can be achieved by clustering the embeddings in the database based on their vector similarity, which identifies thematic similarities among the question-answer pairs. A popular method for performing such clustering is the K-means algorithm – an unsupervised machine learning technique that partitions the data into a pre-specified number of clusters. To reduce the noise and increase commonalities, we focus on the selection of LFS covering Africa (Figure 2).

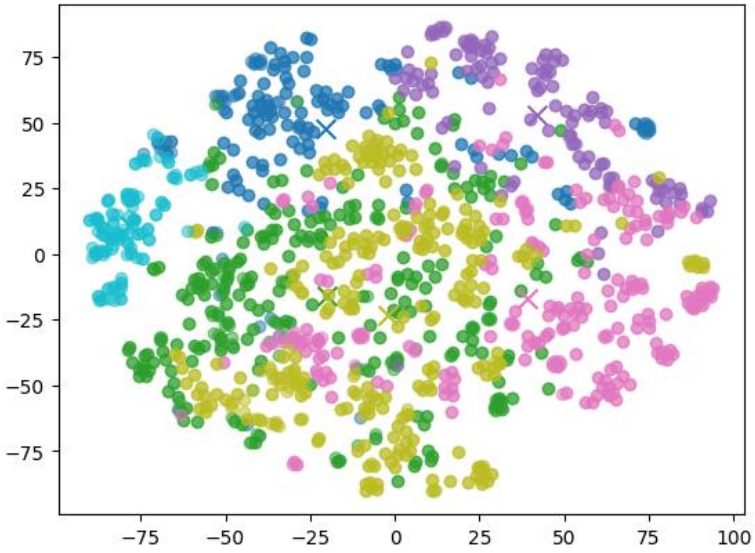
To determine the optimal number of clusters for K-means, we first assess how well different cluster numbers represent the underlying data. Two widely accepted methods – the Elbow Method and the Silhouette Method –guide this decision. The Elbow Method visualizes how clustering improves as we increase the number of clusters by calculating how compact the clusters are (measured as "inertia"). Initially, adding more clusters brings points closer to their centres, but beyond a certain point, the improvements are minimal, forming an "elbow" in the plot and indicating the optimal number of clusters. The Silhouette Method, in contrast, evaluates how distinct and well-separated the clusters are. A higher silhouette score reflects better-defined clusters, with points within each cluster being close together and well-separated from others. By combining these two methods, we determine that six clusters best fit our dataset.

[7] The app is accessible to ILO staff and external researchers, but access requires pre-authorization. Contact Pawel Gmyrek (gmyrek@ilo.org) to obtain access.

▶ **Figure 1. Content of the LFS database: question-answer pairs by country and income group**

Figure 2 provides a visual representation of the clustering exercise. It is important to note that this visualization is limited to two dimensions, whereas the embeddings used consist of 1536 dimensions. This "flattening" of the vector space helps to visualize the most obvious clusters, but more nuanced groupings may be harder to discern visually.

▶ **Figure 2. LFS covering Africa: 2D visual of semantic clustering (t-SNE algorithm)**



Note: The figure shows a 2D visualization of clusters using the t-SNE (t-distributed Stochastic Neighbour Embedding) algorithm. Each colour represents a different cluster, grouping semantically similar data points. t-SNE reduces high-dimensional data into 2D, preserving the relative distances between points for easier visualization. The X-axis and Y-axis represent the two main components of this reduced space, showing the spread of clusters.

At this stage, we can again leverage the capabilities of GenAI models to provide a deeper understanding of the clusters. We randomly select a sample of question-answer pairs from each cluster and pass them through the GPT-4o API. The model is instructed to generate a succinct title and a short description of each cluster based on the commonalities in the sample. This results in the identification of titles and descriptions for the thematic clusters, which can be matched by sample questions, as shown in table 3.

▶ **Table 3. Thematic clusters identifiying common thematic areas in the LFS forms covering Africa**

Cluster	Theme	Summary	Sample Questions
1	**Work Hours, Absence, and Workplace Conditions**	The questions focus on working hours, work absences, income continuation during absence, safety procedures, and union membership in the workplace.	Q9. Pendant combien d'heures désirez-vous travailler par semaine? Q3. OSH_DOC Is there any document that explains responsibilities and procedures on health and safety?
2	**Business Operations and Financial Practices**	The questions focus on various aspects of business operations, including the impact of credit, export activities, industry classification, and financial expenditures.	Q13c. Quel a été l'impact du crédit sur votre entreprise? QP13b. Combien s'élève votre revenu (en FCFA)?
3	**Household Demographics and Living Conditions**	These questions cover various aspects of household composition, living conditions, drinking water access, and demographic information such as household membership.	Q L7. Approvisionnement en eau de boisson A 1- Eau courante à la maison Q23. While absent, is/was (NAME) a member of another household?
4	**Household Economic and Living Conditions**	These questions focus on the economic and living conditions of households, including income sources, cost of housing amenities, and educational expenses.	Q5.11 Est-ce que (Nom) a reçu des revenus mobiliers et financiers? Q (11.03) Le logement dispose-t-il des équipements suivants?

5	**Employment Decisions and Barriers**	The questions cover employment-related decisions, job search behaviours, financial borrowing choices, and conditions for job acceptance or rejection.	Q2. Why did you refuse? (Select the main reason) Q22. Pourquoi avez-vous refusé cette proposition?
6	**Education Level and School Attendance**	These questions focus on an individual's education level, school attendance history, and highest grade completed.	QM20a. Quel niveau d'enseignement avez-vous atteint? Q2.31 Quelle est la dernière classe fréquentée par [NOM]?

As this exercise demonstrates, there is potential for more structured, in-depth analysis of survey questionnaires, for example, to identify similar survey questions that have the potential to be standardized across countries.

Conclusions

In this technical brief, we have provided an overview of a new database of LFS questions available to ILO researchers. In addition, we described the detailed steps leading to the creation of this database, focusing on the unique capabilities and advantages of latest GenAI tools, specifically their ability to structure unstructured text data and enhance the search and retrieval process through semantic understanding. We believe that a similar set of steps could be applied to the vast number of textual resources held by the ILO and other UN institutions, such as general survey forms, observations of the ILO Committee of Experts, minutes of the Governing Body and ILC discussions, or evaluation reports and Development Cooperation project documents.

As demonstrated in our example of LFS forms, digitization of documents can mean dramatically different things. Although many ILO documents are stored digitally in raw formats like .pdf or .xlsx, these formats offer limited possibility for bulk processing of textual information. GenAI-powered tools can transform them into structured, query-ready datasets, allowing for more dynamic and efficient research and policy development.

It should, however, be noted that implementing these technologies requires a certain investment in human effort and technical expertise, particularly in areas such as data science, Natural Language Processing (NLP), and AI tool integration. While familiarity with programming languages like R or Python is beneficial, individuals can often start with basic to intermediate proficiency. More advanced expertise can be built gradually as they become more comfortable with these tools. Once implemented, GenAI tools can help streamline the operational aspects of handling textual resources and contribute to the speed and quality of ILO's analysis and policy advice. Such preparatory effort should be considered an investment for the future, as organized and structured materials become a new digital asset that can be reused as the next generations of AI tools evolve and gain new data processing abilities.

As part of the ILO's institutional knowledge-sharing efforts, we have provided a synthetic version of the Python code used for this exercise on a publicly accessible GitHub repository. We hope this resource will encourage further experimentation and innovation with GenAI tools, allowing researchers and practitioners to advance their technical expertise and extend the benefits of these methods across a range of disciplines.

Contact details
International Labour Organization
Route des Morillons 4
CH-1211 Geneva 22
Switzerland

T: +41 22 799 7239
Pawel Gmyrek
E: gmyrek@ilo.org

DOI: https://doi.org/10.54394/SORG5455