The Greek Review of Social Research, 2024, 163
Print ISSN: 0013-9696 Online ISSN: 2241-8512
Copyright © 2024 The Author(s)
This work is licensed under the Creative Commons
Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
https://creativecommons.org/licenses/by-nc/4.0/

Apostolos Linardis[*], *Dimitra Kondyli*[**], *Konstantinos Alexandris*[***], *Konstantinos Papagiannopoulos*[****], *Konstantinos-Symeon Nisiotis*[*****], *Nikolaos Mastoris*[******], *Nicolas Klironomos*[*******]

CONFIGURING, OPTIMIZING AND ENHANCING DATAVERSE: THE CASE OF SODANET REPOSITORIES

ABSTRACT

Current trends in computational social sciences and data-driven research require access to reliable data sources. Data repositories fulfil this role, providing high-quality data services, since they satisfy certain requirements that can be briefly summarized as complying with the FAIR data principles. In this context, we outline the measures undertaken by SoDaNet to enhance the data repositories within our infrastructure. Through the adoption and implementation of the Dataverse software, along with a series of adaptations and customizations, we have improved the Findability, Accessibility, Interoperability, and Reusability (FAIR) of the data hosted in these repositories. This effort has resulted in data repositories responding effectively to the demands of the research community, improving the user experience both nationally and internationally.

Keywords: *Data Repositories, metadata, FAIR data, data management*

[*] Research Director, National Centre for Social Research, President of the SoDaNet General Assembly, e-mail: alinardis@ekke.gr

[**] Research Director, National Centre for Social Research, President of the SoDaNet Steering Committee, e-mail: dkondyli@ekke.gr

[***] Scientific Associate, National Centre for Social Research, e-mail: kostisalex@gmail.com

[****] Scientific Associate, National Centre for Social Research, e-mail: kpapag@ekke.gr

[*****] Scientific Associate, National Centre for Social Research, e-mail: csnisiotis@yahoo.gr

[******] Scientific Associate, National Centre for Social Research, e-mail: mastorisnick@gmail.com

[*******] Scientific Associate, National Centre for Social Research, e-mail: nklironomos@ekke.gr

Received: April 26, 2023 Accepted: January 17, 2024 Published: July 30, 2024

Απόστολος Λιναρδής[*], *Δήμητρα Κονδύλη*[**], *Κωνσταντίνος Αλεξανδρής*[***], *Κωνσταντίνος Παπαγιαννόπουλος*[****], *Κωνσταντίνος-Συμεών Νησιώτης*[*****], *Νίκος Μάστορης*[******], *Νίκος Κληρονόμος*[*******]

ΠΡΟΣΑΡΜΟΓΗ, ΠΑΡΑΜΕΤΡΟΠΟΙΗΣΗ ΚΑΙ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΤΟΥ DATAVERSE: Η ΠΕΡΙΠΤΩΣΗ ΤΩΝ ΑΠΟΘΕΤΗΡΙΩΝ ΔΕΔΟΜΕΝΩΝ ΤΟΥ SODANET

ΠΕΡΙΛΗΨΗ

Η ανάγκη για πρόσβαση σε αξιόπιστες πηγές δεδομένων κρίνεται πιο επιτακτική από ποτέ, λαμβάνοντας υπόψη τις εξελίξεις στην υπολογιστική κοινωνική επιστήμη και την έρευνα που βασίζεται σε δεδομένα. Τα αποθετήρια δεδομένων ανταποκρίνονται σε αυτήν την ανάγκη, προσφέροντας υπηρεσίες δεδομένων υψηλής ποιότητας, υπό την προϋπόθεση ότι συμμορφώνονται με κριτήρια που αποσκοπούν στην εφαρμογή των αρχών FAIR. Στο πλαίσιο αυτό, περιγράφουμε τις ενέργειες που υλοποιήσαμε ως SoDaNet με στόχο να εξασφαλίσουμε ότι τα αποθετήρια δεδομένων της υποδομής μας, μέσω της υιοθέτησης και της εφαρμογής του λογισμικού Dataverse, καθώς και μέσα από μια σειρά προσαρμογών και παραμετροποιήσεων, έχουν βελτιωθεί ώστε τα δεδομένα που φιλοξενούνται σε αυτά να είναι πιο εύκολα εντοπίσιμα, προσβάσιμα, διαλειτουργικά και επαναχρησιμοποιήσιμα (FAIR). Αυτή η προσπάθεια είχε ως αποτέλεσμα τα αποθετήρια δεδομένων του SoDaNet να ανταποκρίνονται αποτελεσματικά στις απαιτήσεις της ερευνητικής κοινότητας, ενισχύοντας την εμπειρία του χρήστη, τόσο σε εθνικό όσο και σε διεθνές επίπεδο.

Λέξεις κλειδιά: Αποθετήρια Δεδομένων, μεταδεδομένα, FAIR δεδομένα, διαχείριση δεδομένων

[*] Διευθυντής Ερευνών, ΕΚΚΕ, Πρόεδρος της Γενικής Συνέλευσης του SoDaNet, e-mail: alinardis@ekke.gr

[**] Διευθύντρια Ερευνών, ΕΚΚΕ, Πρόεδρος της Διοικούσας Επιτροπής του SoDaNet, e-mail: dkondyli@ekke.gr

[***] Επιστημονικός συνεργάτης, ΕΚΚΕ, e-mail: kostisalex@gmail.com

[****] Επιστημονικός συνεργάτης, ΕΚΚΕ, e mail: kpapag@ekke.gr

[*****] Επιστημονικός συνεργάτης, ΕΚΚΕ, e-mail: csnisiotis@yahoo.gr

[******] Επιστημονικός συνεργάτης, ΕΚΚΕ, e mail: mastorisnick@gmail.com

[*******] Επιστημονικός συνεργάτης, ΕΚΚΕ, e mail: nklironomos@ekke.gr

INTRODUCTION

SoDaNet[1] is one of the 28 national research infrastructures in Greece but the only one in the field of social sciences. Sodanet provides many services to its members and the wider academic and research community, including access to data through a data catalogue, access to courses on data management and research methodology, long-term preservation of third-party data deposited in the data catalogue repositories, consultancy services for the creation of Data Management Plans (DMPs), training services, online survey services and IT services.

The networking of SoDaNet aims to spread at a national and European level. The national research network was established in 2012 and consists of the National Centre for Social Research of Greece and 7 university departments of social sciences[2]. In June 2015, SoDaNet became a member of the Consortium of European Social Science Data Archives (CESSDA) - ERIC and since then it has been operating as the Greek hub of the European infrastructure.

The primary SoDaNet service is to provide access to data and metadata for secondary use and analysis through its data catalogue. In 2019, SoDaNet decided to create a new data catalogue, aiming to replace the current one, as a move towards enhancing and streamlining the repository services provided. The new SoDaNet Data Catalogue[3] was implemented in the context of the national project titled “SoDaNet in Action”, funded by the Operational Programme for Competitiveness, Entrepreneurship and Innovation 2014-2020 (EPAnEK) and the European Regional Development Fund (ERDF).

The former data catalogue was based on the Nesstar software. Although the Nesstar software proficiently handled quantitative datasets, including the documentation of variables and questions, and facilitated online statistical analysis and geographic visualization, it was disadvantaged by the lack of ongoing maintenance or development. Thus, lacked up-to-date features necessary in a state-of-the-art repository software, made its use as a basic repository software questionable and led SoDaNet to seek reliable solutions. A software solution endorsed by SoDaNet was the Dataverse application. Dataverse is an open-source application for sharing,

[1] The SoDaNet website is available here: https://sodanet.gr/

[2] The institutions forming SoDaNet are listed here: https://sodanet.gr/about/sodanet-members

[3] The SoDanet Data Catalogue is available here: https://datacatalogue.sodanet.gr

discovering and preserving data developed by the Institute for Quantitative Social Sciences (IQSS) at Harvard University and subsequently embraced by a growing community of users and adopters worldwide. It was initially developed to solve problems related to the dissemination of social data, such as the need to identify the authors of the data, their long-term preservation as well as their public and open publication (King 2007; Crosas 2011). Dataverse as a repository software, offered a range of desirable features (i.e., Persistent Identifiers (PIDs), User Management, Folder Management, Data Documentation Initiative (DDI) compliance, File Management etc.) whilst lacking some of the robust features that were available on Nesstar. However, being an open-source software implies that it could be conveniently adapted to meet the repository needs of the infrastructure. So, the Dataverse ver. 4.16 solution has been adopted based on the strategic decision that the required configurations would be implemented. The requirements of the repository were addressed along the following two axes:

- Domestic requirements. Given that SoDaNet is a consortium of 8 members, the new data catalogue needs to be functional while considering the unique aspects of each partner (i.e., partner expertise in quantitative/qualitative research data) and also fulfilling national requirements for data documentation and management (Linardis & Ioannidis, 2022).
- Non-domestic requirements. SoDaNet is a Service Provider (SP) of the CESSDA ERIC, therefore some specific requirements and standards had to be fulfilled (i.e., compliance with the Data Documentation Initiative[4] and with CESSDA Controlled Vocabularies[5] - SoDaNet data should be accessible via the CESSDA Data Catalogue).[6]

Simultaneously, strategic guidelines at both the European and national levels, as well as international best practices for data, demand compliance with the principles of Findable, Accessible, Interoperable, and Reusable

[4] The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences. https://ddialliance.org/

[5] CESSDA Controlled Vocabulary Service: https://vocabularies.cessda.eu/

[6] The Cessda Data Catalogue (CDC) is a multilingual catalogue that provides information (metadata) on more than 40,000 datasets, available at https://www.cessda.eu/Tools/Data-Catalogue

(FAIR) data and Open Science (Kondyli & Klironomos, 2022; GOFAIR, n.d). The overarching goal is to support the secondary use/reuse of scientific data (Kondyli, Nisiotis & Klironomos, 2024).

CONFIGURING, OPTIMISING & ENHANCING DATAVERSE

In the following sections, we will examine in detail the operations performed by SoDaNet to customise, optimise and enhance the Dataverse software to fulfil the current needs of the research community.

DATA PROJECTS & RESOURCES

One of the requirements of the users and members of the Greek network was to distinguish between the datasets hosted in the infrastructure.

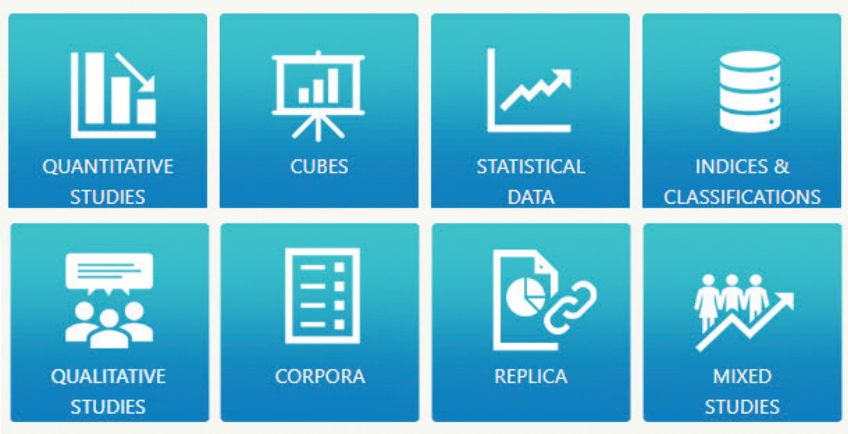


Figure 1: *The of data projects supported by SoDaNet Dataverse after its expansion*

This distinction resulted in the following eight objects (see Figure 1):

- Quantitative studies.
- Qualitative studies.
- Mixed studies.
- Statistical Research Metadata.
- Cubes.
- Indicators & Classifications.

- Replicas for the reproduction of statistical analysis, and
- Corpora.

Those eight objects do not necessarily refer to datasets but also to “metadata sets” (e.g. Indicators and Classifications, Statistical Research Metadata). For this reason, the default term “datasets” -adopted by Dataverse- has been replaced by the broader term “data projects”, i.e., data projects that can refer to either projects that contain data or projects that contain metadata (Linardis & Ioannidis 2022; Kallas & Paparisteidi, 2022). Moreover, we’ve substituted the technical term “file” (commonly used in Dataverse) with the more universally recognized term “resource”. The broader concept of a “resource” encompasses not only data but also tools, like questionnaires and codebooks, as well as essential metadata (Linardis, Maravelakis & Fragoulis, 2023). Notably, a data project can consist of zero (in the case of “metadata sets”), one, or multiple resources.

In addition to the quantitative, qualitative and mixed studies that concern studies mentioned in most books that deal with research methodology, it is appropriate at this point to briefly refer to the other five data projects. A cube is a multidimensional table containing aggregated data and is structured by one or more variables expressing dimensions and at least one variable / statistical function expressing a measure. An indicator or a classification is a target variable that is derived from the processing of one or more source variables through an applied sequence of logical and arithmetic operations (algorithm). The replica consists of the algorithm and the subset of data used to answer the research questions reported in a publication. The statistical research metadata concerns the documentation of surveys carried out by the Hellenic Statistical Authority or from institutions of the Hellenic Statistical System. These institutions document their surveys according to the SIMS (Single Integrated Metadata Structure) standard, while they submit their data to Eurostat (Hellenic Statistical Authority, 2022). Finally, corpora concern data resources consisting of one or more large and structured sets of texts.

Extension of the Data Project metadata schema

To meet documentation requirements, the default metadata schema of Dataverse was expanded and customized. This was done for the following reasons:

- the eight Data Project categories (Figure 1) were added as a metadata field in the Dataverse “Citation Metadata” module,

- several new fields were added to the existing metadata modules to support full documentation of quantitative, qualitative, and mixed studies as well as replicates, cubes and corpora,
- two distinct new metadata modules were created to document:
  - Indicators & Classifications, and
  - Statistical Research Metadata according to SIMS Metadata Structure.

In total, more than 150 new metadata fields were added. Most of them are concerned with the documentation of Statistical Surveys according to SIMS (Hellenic Statistical Authority, 2022).

The types of metadata fields added/changed were:

- Free-text fields (either plain text or HTML).
- Controlled Vocabularies (CVs), i.e., lists of values with single or multiple selections. Those CVs are based on the CESSDA CVs (e.g. Analysis Unit, Topic Classification, Mode of Collection etc.) and were set in an existing or in a new metadata field.
- Complex metadata fields comprise more than one subfield. Complex metadata fields are metadata fields that belong to a more general grouping. This field has multiplicity, providing the ability to add as many subfields as needed.

To extend the metadata schema of the data project, the native Dataverse mechanism was utilised and extended by uploading TSV (Tab-Separated Values) files.[7] Apart from the extension of the metadata at the data project level, it was considered crucial to extend also the resource metadata.

Extension of the resource metadata schema

The interface for editing/documenting resources has been enhanced to incorporate a range of functionalities, distributed across five distinct modules, as outlined below:

- General Resource Information.
- Variable description.
- Variable groups.
- Questions.
- Cube layout.

[7] More information on Dataverse metadata customization is available here: https://guides.dataverse.org/en/latest/admin/metadatacustomization.html

It’s important to highlight that in Dataverse version 4.16, the information available at the resource level was limited to several “technical” fields regarding the “General Resource Information” module. Structural entities like variables, questions, and cubes were not distinctly recognized by the original Dataverse system for tabular files that were ingested in the database (Tabular files typically consist of rows and columns, where each row represents a record and each column represents a specific attribute or variable of the data). The documentation of variables was deemed necessary for the secondary use of the tabular files. The demand from SoDaNet members for a catalogue capable of effectively managing quantitative surveys and data necessitated the inclusion of documentation for variables and questions for tabular files. Tabular files may contain aggregated data, known as cubes, or “microdat’’ that originate from surveys conducted via recording forms or questionnaires. For both microdata and cubes, it is essential to provide metadata about the “General Resource Information’’, “Variable Description’’, and “Variable Groups’’. Metadata for “Cube layout” should be used for cube documentation, while metadata for questions should be used for microdata. It is essential to document the original questions that variables stem from, allowing secondary users to comprehend the exact wording used to derive these variables. Simultaneously, this enhancement ambitiously seeks to engage with initiatives like the “CESSDA EuroQuestionBank (EQB)’’, which strives to establish a cross-national question bank. This bank would feature a central search capability throughout all CESSDA survey collections, presenting survey questions in multiple languages through an intuitive application (Akdeniz & Zenk-Möltgen, 2019).

Incorporating these modules greatly assists users in understanding survey-derived tabular resources by providing full provenance details. This capability allows users to evaluate the resource’s usefulness, thus promoting the reutilization of survey data within the infrastructure.

General Information on Resources

As Dataverse offers few metadata fields at the resource level, the resource management interface was redesigned and implemented from scratch. Ten (10) new metadata fields were added at the resource level for a total of thirteen (13) fields (free-text and CVs). Since there is no native Dataverse mechanism for resource management, both the software and the database schema had to be extended accordingly. The ten new fields that have been added are:

- Resource title (Dataverse did not include the resource title field but used to store and show only the filename of a resource instead. The whole new user interface was redesigned to expose primarily the Resource Title and secondarily the filename).
- Language (CV).
- DOI/URL.
- Resource type (CV).
- Notes.
- Data source.
- Access upon login (Yes/No).
- Embargo period (date field indicating the end date of the embargo period for the resource in question).
- License (Creative Commons Licenses such as: CC BY, CC BY–SA, CC BY-ND, CC BY-NC, CC BY-NC-SA, CC BY-NC-ND).
- Visualization hints (for the needs of pivoting to cubes).

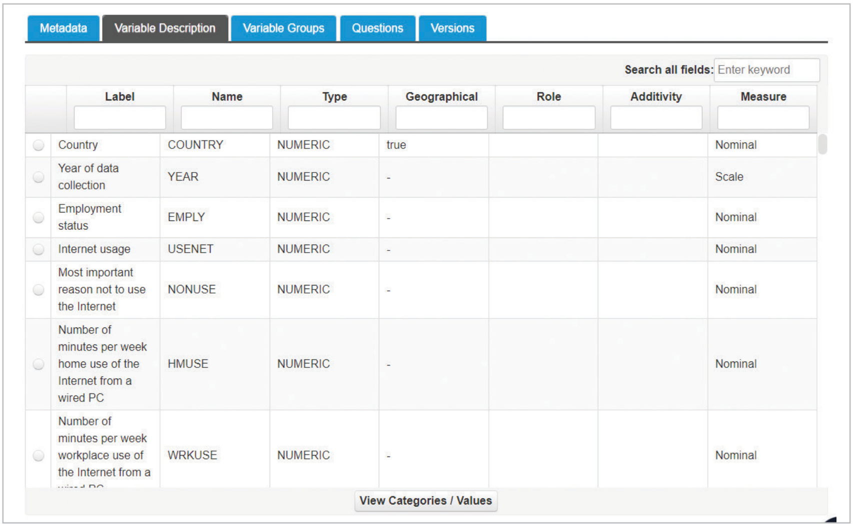
Variable Description

This feature provides the ability to edit and manage the variables of a resource, which Dataverse inherently understands as "tabular". Specifically, Dataverse's upload, edit, and ingest mechanism has been extended to allow access and editing of the list of variables in a tabular file. The database has also been extended so that the description of a variable can be stored with the following properties (Figure 2):

- Label.
- Name.
- Type (selection from CV).
- Geographic (selection from CV: this refers to the link to specific maps provided by the SoDaNet cartographic application).
- Role (option from CV: indicates whether it is a weighting or identification variable).
- Additivity (stock, flow, non-additive). A property mainly related to the management of cubes.
- Measurement (selection of CV, such as: nominal, ordinal, scale).
- Categories / Values: definition of codes and categories of variables with the possibility of defining missing values.

Each of the properties above is a new column in the data-variable table of the database schema. To facilitate the management of variables, the variable list table provides the following functions (Figure 2):

- Variable search by entering a keyword.
- Filtering variables by properties.
- Layout with the option to set the number of variables displayed per page.



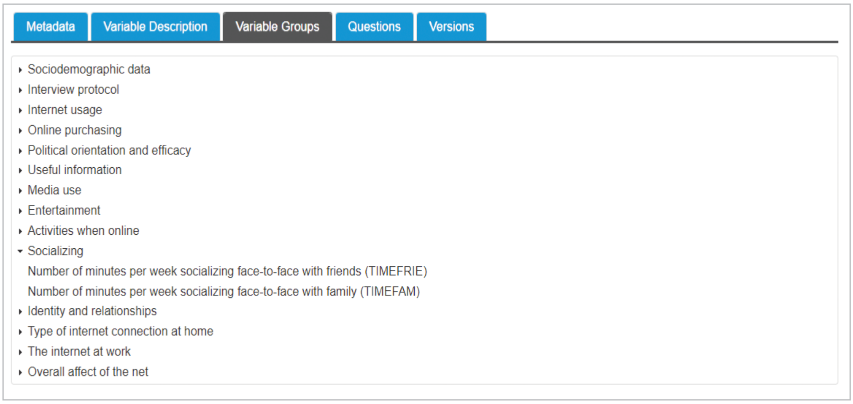
	Label	Name	Type	Geographical	Role	Additivity	Measure
○	Country	COUNTRY	NUMERIC	true			Nominal
○	Year of data collection	YEAR	NUMERIC	-			Scale
○	Employment status	EMPLY	NUMERIC	-			Nominal
○	Internet usage	USENET	NUMERIC	-			Nominal
○	Most important reason not to use the Internet	NONUSE	NUMERIC	-			Nominal
○	Number of minutes per week home use of the Internet from a wired PC	HMUSE	NUMERIC	-			Nominal
○	Number of minutes per week workplace use of the Internet from a wired PC	WRKUSE	NUMERIC	-			Nominal

Figure 2: *The description of the variables in the frontend of the application's extension*

Variable groups

Another extension within tabular files is the ability to add/edit/delete variable groups. A variable group consists of the following structural elements (Figure 3):

- Code: unique identifier for each group.
- Name of the variable group.
- Description: a brief description of the group.
- Variables: the set of variables that belong to this group. The set of variables is available through a multiple-choice list format with filtering (variable name, variable description). The variables are dynamically retrieved from the database as they are stored during the ingestion process, after the resource has been uploaded.

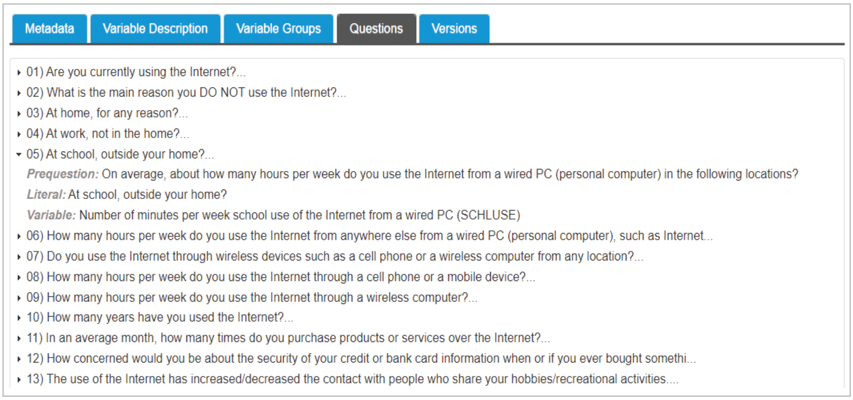


- Sociodemographic data
- Interview protocol
- Internet usage
- Online purchasing
- Political orientation and efficacy
- Useful information
- Media use
- Entertainment
- Activities when online
- Socializing
  Number of minutes per week socializing face-to-face with friends (TIMEFRIE)
  Number of minutes per week socializing face-to-face with family (TIMEFAM)
- Identity and relationships
- Type of internet connection at home
- The internet at work
- Overall affect of the net

Figure 3: *The definition of variable groups in the frontend of the application's extension*

Questions

Furthermore, within the Questions module, there's the capability to associate the variables of a tabular file with the questions they originate from. Each question consists of six basic properties (Figure 4):



- 01) Are you currently using the Internet?...
- 02) What is the main reason you DO NOT use the Internet?...
- 03) At home, for any reason?...
- 04) At work, not in the home?...
- 05) At school, outside your home?...
  Prequestion: On average, about how many hours per week do you use the Internet from a wired PC (personal computer) in the following locations?
  Literal: At school, outside your home?
  Variable: Number of minutes per week school use of the Internet from a wired PC (SCHLUSE)
- 06) How many hours per week do you use the Internet from anywhere else from a wired PC (personal computer), such as Internet...
- 07) Do you use the Internet through wireless devices such as a cell phone or a wireless computer from any location?...
- 08) How many hours per week do you use the Internet through a cell phone or a mobile device?...
- 09) How many hours per week do you use the Internet through a wireless computer?...
- 10) How many years have you used the Internet?...
- 11) In an average month, how many times do you purchase products or services over the Internet?...
- 12) How concerned would you be about the security of your credit or bank card information when or if you ever bought somethi...
- 13) The use of the Internet has increased/decreased the contact with people who share your hobbies/recreational activities....

Figure 4: *Defining the questions in the front-end of the application's extension*

- Code: unique identifier for each question.
- Pre-question: free-text field that refers to the existing text before the literal question.
- Literal question.
- Post question: free text that refers to the text after the literal question.
- Instructions are given for the correct completion of the question.

- The variable that was derived from the question. Only one variable can be selected for each question (1 to 1 relationship between question and variable).

Cube Layout

For data that contain variables but include aggregated data, the documentation has been extended to specify the initial layout of a resource-level cube by specifying the rows, the columns, and the measures of the table. It should be noted that the variables are selected from the variable pool which is created after the data file is uploaded, processed and ingested into the database. Afterwards, it is possible to display the cube using the appropriate pivoting tool.

MAKING DATA FINDABLE

The enhancements mentioned above, whether expanding data project metadata or resource metadata, were designed to improve the discoverability and identification of data projects and resources. In line with the GOFAIR principles (n.d), it is emphasized that metadata and data should be easily accessible to both humans and computers. Metadata that is readable by machines plays a crucial role in the automated discovery of datasets and services. To ensure the SoDaNet repositories serve its users effectively, especially considering the predominance of Greek-speaking users, enhancing the ability for users to locate desired data was deemed critical. Therefore, efforts were made to embody the GOFAIR principles of assigning a globally unique and persistent identifier to metadata/data, enriching data descriptions with comprehensive metadata, and making metadata/data searchable within a resource. The sections below detail the implementation of Digital Object Identifiers (DOIs) as persistent identifiers (PIDs) and the enhancement of the search functionality.

DOIs assigned for each Data Project

The operational need to connect the data catalogue to the production and distribution platform for DOIs led to the compelling need to configure and extend Dataverse to:

- Utilize the Dataverse API to transmit the relevant metadata to the Datacite platform efficiently.

- Enable the creation of both dummy and authentic DOIs within the Datacite test environment for the staging setup. For the production setup, facilitate the generation of genuine DOIs on the Datacite production platform.
- Automate the assignment of DOIs through a seamless authentication process on the Datacite platform, leveraging the Dataverse API.

Improved search mechanism

Dataverse version 4.16, which serves as the basis for the SoDaNet repository, utilizes Solr for its core indexing and search functionalities, covering both data and metadata. Solr, renowned as a leading open-source search engine developed in Java, is widely acclaimed in its field. Nonetheless, it's important to note that Solr version 7.3, currently in use, is affected by a critical security vulnerability. This issue underscores the need for an upgrade to the latest version, ver. 8.9, to bolster security and maintain robust search capabilities.

The upgrade of Solr has brought about several significant improvements, including:

- Resolution of security vulnerabilities found in earlier versions.
- Enhancement of the search engine's response times.
- Support for the incorporation of sophisticated search algorithms, such as:
  - Intonation-based word search.
  - Stemming for more accurate search results.
  - Advanced search capabilities specifically designed for Greek language entries, accommodating grammatical nuances like gender, cases, and plural forms, among others.

The above search enhancements were made to cover both the metadata fields of data projects, resources and variables.

MAKING DATA ACCESSIBLE

When a user locates the necessary data, it's crucial to provide information on how to access it, which may involve steps for authentication and authorization (GOFAIR, n.d.). We decided that at least for any resource, where further possibilities for visualizations are given, we should ask for user registration and authentication (i.e. the "Explore upon login" button). So, we extended Dataverse to "understand" the "Access upon login" access level. We also implemented the embargo period feature for each resource, that is a date field that specifies the end date of the embargo period. After the implementation of the aforementioned extensions, and in combination with the inherent Dataverse's native access levels, the following four (4) access levels are supported per resource:

- Unrestricted access
- Access upon login
- Restricted - Available on request
- Universal restricted

The user can set a combined embargo period for the first three access categories. Additionally, anyone with a computer and an internet connection can access at least the metadata through the HTTPS open and free (no-cost) protocol, globally implementable to facilitate data retrieval (GOFAIR, n.d.). Below are given details for the implementation of the "Access upon login" access level and the embargo period per resource.

Access upon login

This feature is designed to allow access to certain resources only to logged-in users, i.e., users who have the appropriate permissions in the data catalogue via the Sodanet Single-Sign-On (SSO) user authentication system. To restrict access to a particular resource, the individual responsible for documenting it must choose the "Access upon login" option on the resource metadata editing interface. This action limits access solely to users who are logged in, effectively "locking" the resource for anyone not authenticated (anonymous users, Figure 5).

Embargo[8]

The embargo date functionality is crafted to regulate the timing of when published resources are made available for detailed viewing (such as variable displays) or downloading, distinct from their actual date of publication. This allows the publisher to upload the resource at their convenience and specify an embargo date, making the resource accessible to end users only once this date has passed. Implementation of the embargo date is facilitated through a dedicated date field within the resource metadata

[8] It should be noted that this feature was implemented by SoDaNet in Dataverse version v4.16 in August 2019, while the Dataverse community implemented it in version 5.8 in November 2021. Dataverse 5.8 release notes: https://dataverse.org/blog/dataverse-software-58-release

form. To accommodate this functionality, both the database and the source code have been updated and expanded, ensuring they can effectively store and manage this information.

MAKING DATA INTEROPERABLE

For data to be interoperable, it must seamlessly interact with applications or workflows dedicated to analysis, storage, and processing (Wilkinson et al., 2016; GOFAIR, n.d.). As previously noted, we employed CESSDA CVs to guarantee metadata field interoperability. Additionally, measures were implemented to facilitate the harvesting of data, ensuring its broad availability in the CESSDA Data Catalogue. From there, the data can be further disseminated to larger research data aggregators, including OpenAIRE.[9] The Dataverse installation makes local data project metadata available to remote harvesting clients through the OAI-PMH protocol.[10] Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability, where repositories expose structured metadata.[11] Dataverse by default is compliant with citation and domain-specific metadata such as DDI Lite, DDI 2.5[12] and Dublin Core and exposes metadata through OAI-PMH. Nevertheless, in practice, the installation of Dataverse performs a lot of schema violations, concerning the compliance with DDI 2.5, which is one of the profiles to be compliant with the CESSDA Data Catalogue. CESSDA provides a metadata validation tool[13] to Service Providers, to test compliance with specific profiles such as DDI 1.2.2, DDI 2.5 and DDI 3.2. The next actions were taken for SoDaNet's Data Catalogue to be included in the CESSDA Data Catalogue endpoint:

- A tool was developed to validate all data projects using a specific XSLT. This tool evaluates the generated XML files and compiles a

[9] CESSDA Data Catalogue is visible through the OpenAIRE EXPLORE at the followinglink: https://explore.openaire.eu/search/dataprovider?datasourceId=re3data::bcf017a6702b1ff5fcaccca1ca44c010

[10] Instructions for setting up a Dataverse Installation as an OAI server can be found here: https://guides.dataverse.org/en/latest/admin/harvestserver.html

[11] More information on OAI-PMH protocol can be found here: http://www.openarchives.org/pmh/

[12] More information on Dataverse's supported metadata standards is available here: https://guides.dataverse.org/en/latest/user/appendix.html

[13] CESSDA validation tool is available here: https://cmv.cessda.eu/

comprehensive report detailing any discrepancies. The configuration of this mechanism is flexible, allowing for various XSLTs to be used as input and generating corresponding reports. It will soon be accessible as a web-based service.

- Following the insights from this report, we made strategic updates to Dataverse to align with the CESSDA Data Catalogue's requirements. Necessary modifications and enhancements were implemented in Dataverse, particularly in the generation of OAI-PMH XML, to ensure compliance. A significant issue addressed was the absence of language specification, which was a prevalent compatibility issue with DDI 2.5 standards.

It should be noted that at the present stage, most of the CESSDA Service Providers that are using Dataverse, are facing problems to be included in the CESSDA Data Catalogue endpoint, due to schema violations of the standard Dataverse installation.

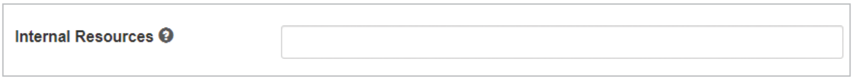
Finally, the addition of numerous new metadata fields made it necessary to improve and extend the XML creation mechanism, both in terms of the OAI-PMH protocol and the native mechanism for creating and exporting XML files based on the DDI schema.

MAKING DATA REUSABLE

To achieve reusability, data are richly described with metadata that meets domain-relevant community standards such as the DDI. Additionally, for each resource, we adopted a clear and accessible data usage license through Creative Commons (GOFAIR, n.d.). Sodanet facilitates open access to resources by trying to persuade depositors to provide licenses that promote open access when there are no use restrictions. In addition to the above, SoDaNet has implemented mechanisms for the reuse of internal (within SoDaNet repositories) and external (existing outside of the SoDaNet repositories) resources. The initial installation of the Dataverse software has no provision for resource reuse. It was necessary to design and develop mechanisms to create storage and view associations for either internal or external resources. In both cases, the database schema, the environment for documenting the data project, and the environment for viewing the metadata of the data project were extended. In both of the above cases, the database structure, the data project documentation environment, and the data project metadata viewing environment were extended.

Making Internal Resources Reusable

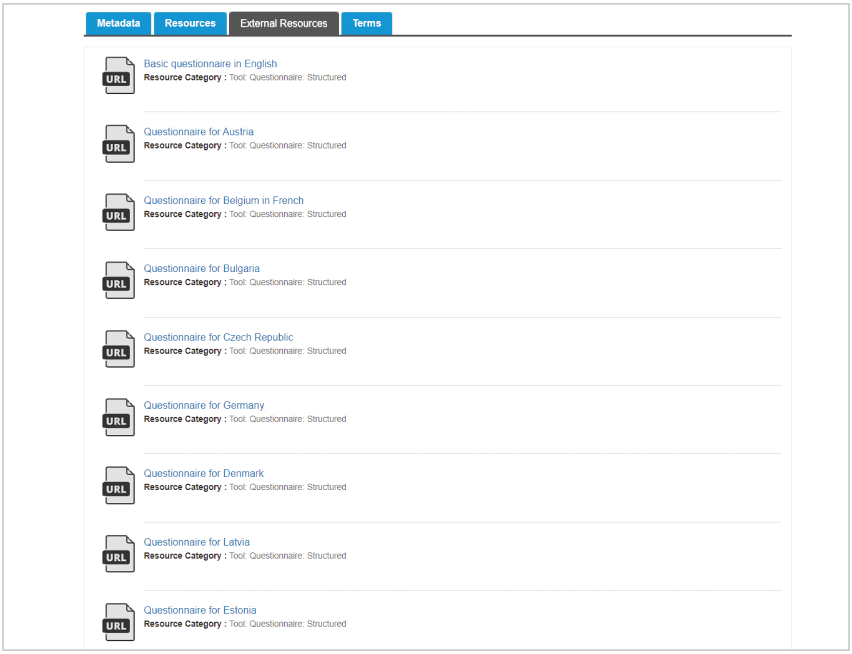
A new field has been introduced within the “Citation Metadata” module to facilitate the reuse of internal resources, enabling users to input multiple codes associated with internal resources (Comma Separated Resource IDs) (Figure 6).



Internal Resources

Figure 6: *Defining internal resource codes for reuse*

To find the identification code of each resource, the user must consult the “Technical Metadata” tab of the resource and copy the value of the “DataCatalogue ID” field. Following the saving of the internal resource interface, the metadata environment of the data project is instantly updated. Consequently, the “Resources tab” will display the linked internal resources. The associated internal resources have the same access rights as those set in the data project that were originally published.



Basic questionnaire in English
Resource Category : Tool: Questionnaire: Structured

Questionnaire for Austria
Resource Category : Tool: Questionnaire: Structured

Questionnaire for Belgium in French
Resource Category : Tool: Questionnaire: Structured

Questionnaire for Bulgaria
Resource Category : Tool: Questionnaire: Structured

Questionnaire for Czech Republic
Resource Category : Tool: Questionnaire: Structured

Questionnaire for Germany
Resource Category : Tool: Questionnaire: Structured

Questionnaire for Denmark
Resource Category : Tool: Questionnaire: Structured

Questionnaire for Latvia
Resource Category : Tool: Questionnaire: Structured

Questionnaire for Estonia
Resource Category : Tool: Questionnaire: Structured

Figure 7: *External resources from the German repository GESIS that are used for the documentation of the Greek translation of the Flash Eurobarometer 241*

Making External Resources Reusable

A new metadata module titled “External Resources” was created to implement associations with external resources. This category includes a complex field with five (5) subfields (Title, Description, Notes, URL and Resource Category). This complex field is implemented with multiplicity, allowing the user to enter more than one external resource. On the data project metadata page, a new tab titled: “*External Resources*” is presented, whenever there is at least one external resource (Figure 7).

MAKING DATA DISCOVERABLE

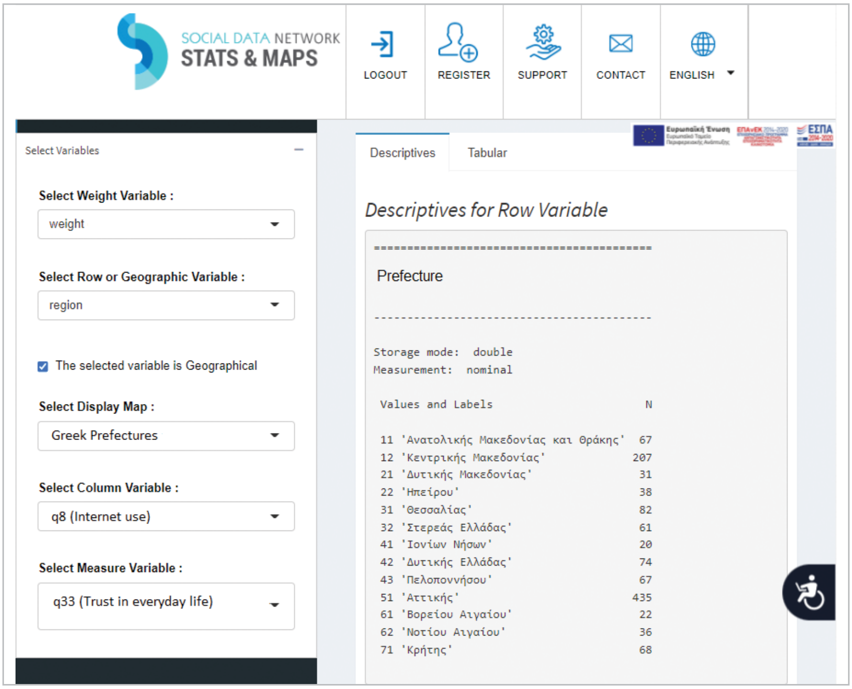
Extending Dataverse to understand and manage structural elements as “variables” gave us additional possibilities for further data discovery. For this reason, we implemented the ‘‘stats’’ application for online generation of descriptive survey microdata statistics, the ‘‘pivoting’’ application for online visualization and management of cube-derived data, and the ‘‘maps’’ application, which maps data involving geographic variables derived from either microdata or cubes.

Online statistics & cartographic visualisations

An application called “Stats & Maps” was implemented using R-Server, R-Shiny and GeoServer for creating online statistical analyses and cartographic visualizations. This application offers the ability to perform online statistical analysis on data files that contain variables, in addition to generating graphical representations. (Figure 8).

For variables that are geographic and have been previously associated with preloaded maps, online cartographic representations are available (Figure 9). Thus, giving the capability of downloading maps locally as well as the associated data. ‘‘Stats & Maps’’ has the versatility to either interface with Dataverse to obtain the necessary variables and information or operate independently as a standalone application.[14] Under this scenario, datasets can be uploaded directly into the application, enabling the creation of online statistical analyses and map-based visualizations.

[14] “Stats and Maps” standalone application is accessible here: https://statsandmaps.sodanet.gr/



Select Weight Variable : weight

Select Row or Geographic Variable : region

☑ The selected variable is Geographical

Select Display Map : Greek Prefectures

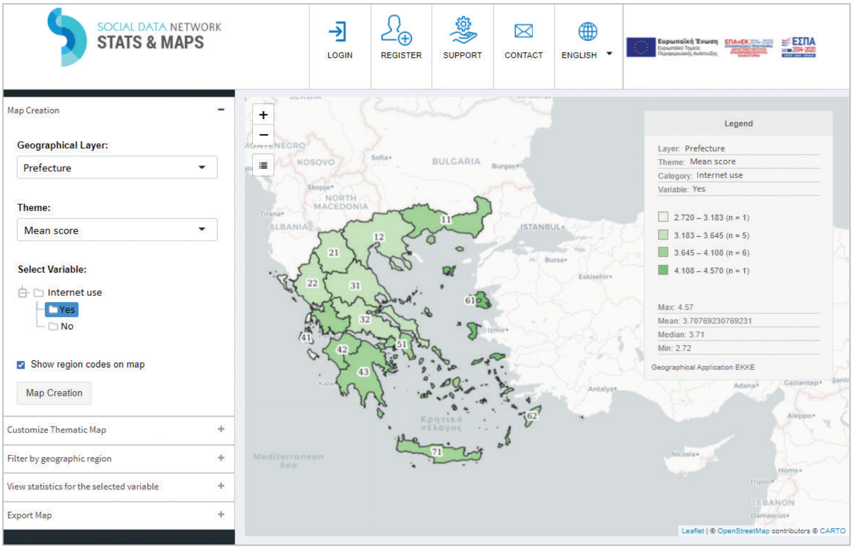
Select Column Variable : q8 (Internet use)

Select Measure Variable : q33 (Trust in everyday life)

Descriptives for Row Variable

```
========================================

Prefecture

----------------------------------------

Storage mode:  double
Measurement:  nominal

 Values and Labels                      N

 11 'Ανατολικής Μακεδονίας και Θράκης'  67
 12 'Κεντρικής Μακεδονίας'             207
 21 'Δυτικής Μακεδονίας'                31
 22 'Ηπείρου'                           38
 31 'Θεσσαλίας'                         82
 32 'Στερεάς Ελλάδας'                   61
 41 'Ιονίων Νήσων'                      20
 42 'Δυτικής Ελλάδας'                   74
 43 'Πελοποννήσου'                      67
 51 'Αττικής'                          435
 61 'Βορείου Αιγαίου'                   22
 62 'Νοτίου Αιγαίου'                    36
 71 'Κρήτης'                            68
```

Figure 8: *StatsAndMaps: an online application for creating descriptive statistics combining up to three variables*



Geographical Layer: Prefecture

Theme: Mean score

Select Variable:
- Internet use
  - Yes
  - No

☑ Show region codes on map

Map Creation

Customize Thematic Map +

Filter by geographic region +

View statistics for the selected variable +

Export Map +

Legend

Layer: Prefecture
Theme: Mean score
Category: Internet use
Variable: Yes

□ 2.720 – 3.183 (n = 1)
🟩 3.183 – 3.645 (n = 5)
🟩 3.645 – 4.108 (n = 6)
🟩 4.108 – 4.570 (n = 1)

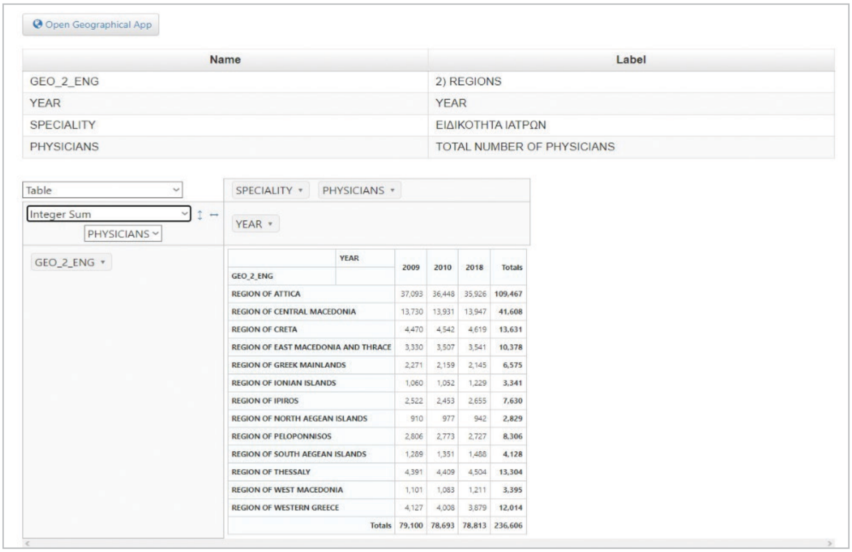
Max: 4.57
Mean: 3.70769230769231
Median: 3.71
Min: 2.72

Geographical Application EKKE

Figure 9: *Stats & Maps: Cartographic representations of variables*

Pivoting application

Especially for the data project category called “Cubes”, an online pivoting application was developed. This tool facilitates the pivoting of the Cube layout that has already been set by the user through the documentation process. That allows the end-users to manipulate the layout of the cube, modify it and generate the preferred table and layout (Figure 10). If the cube includes a geographic variable associated with preloaded maps, it becomes possible to display the generated table visually on a map.



Name	Label
GEO_2_ENG	2) REGIONS
YEAR	YEAR
SPECIALITY	ΕΙΔΙΚΟΤΗΤΑ ΙΑΤΡΩΝ
PHYSICIANS	TOTAL NUMBER OF PHYSICIANS

YEAR	2009	2010	2018	Totals
GEO_2_ENG				
REGION OF ATTICA	37,093	36,448	35,926	109,467
REGION OF CENTRAL MACEDONIA	13,730	13,931	13,947	41,608
REGION OF CRETA	4,470	4,542	4,619	13,631
REGION OF EAST MACEDONIA AND THRACE	3,330	3,507	3,541	10,378
REGION OF GREEK MAINLANDS	2,271	2,159	2,145	6,575
REGION OF IONIAN ISLANDS	1,060	1,052	1,229	3,341
REGION OF IPIROS	2,522	2,453	2,655	7,630
REGION OF NORTH AEGEAN ISLANDS	910	977	942	2,829
REGION OF PELOPONNISOS	2,806	2,773	2,727	8,306
REGION OF SOUTH AEGEAN ISLANDS	1,289	1,351	1,488	4,128
REGION OF THESSALY	4,391	4,409	4,504	13,304
REGION OF WEST MACEDONIA	1,101	1,083	1,211	3,395
REGION OF WESTERN GREECE	4,127	4,008	3,879	12,014
Totals	79,100	78,693	78,813	236,606

Figure 10: *Pivoting of Cube with a geographic variable*

MAKING DATAVERSE MORE USER - FRIENDLY

The user interface underwent a comprehensive overhaul, both in terms of visual appearance and functionality, to enhance the user experience. As part of this effort, the front-end framework was significantly improved with Bootstrap and jQuery technologies, enabling it to support the following features:

- Responsiveness
- Accessibility
- Awareness of the use of cookies
- Multilingualism
- Usability

To support multilingualism, Dataverse has a built-in mechanism that uses property files. Dataverse supports multilingualism broadly, but not in Greek. For this reason, two main actions have been taken:

- Translation of the English user interface into Greek via the Bundle. properties file.
- Augmenting the multilingual capabilities of the open-source software to allow for the translation of messages and elements, like search paths, which were initially supported only in English.

CONCLUSIONS

In summary, the key decision in implementing the SoDaNet data catalogue was to focus on configuring, optimizing, and enhancing Dataverse functionalities. The enhancements, customizations, and improvements made were related to:

- Incorporating Nesstar-like features for managing and analyzing variables, variable groups, and questions.
- Introducing the capability to reuse both internal and external resources in the documentation of a data project.
- Customizing and enhancing native Dataverse mechanisms, including OAI-PMH, persistent DOI assignment, addition and modification of metadata and controlled vocabularies, and the creation of a multilingual user interface.

We think that by integrating these features, SoDaNet repositories have been equipped to meet both domestic and international needs, ensuring that the provided data are as FAIR (Findable, Accessible, Interoperable, Reusable) as possible.

REFERENCES

Akdeniz, E., & Zenk-Möltgen, W. (2019). Metadata Schema of the CESSDA EuroQuestionBank: Documentation and Publication of Survey Questions in a European Question Bank; Version 1.0 (Vol. 2019/15). GESIS - Leibniz-Institut für Sozialwissenschaften. https://doi.org/10.21241/ssoar.65593

Crosas, M. (2011). The dataverse network®: An open-source application for sharing, discovering and preserving data. *D-Lib Magazine*, *17*(1-2) https://10.1045/january2011-crosas

Hellenic Statistical Authority. (2022). Single Integrated Metadata Structure (SIMS) of Labour Force Survey. https://www.statistics.gr/en/statistics/-/publication/SJO01/2022-Q3

GOFAIR (n.d). Fair Principles. https://www.go-fair.org/fair-principles/

Kallas, J., & Paparisteidi, M. (2022). The particularities of empirical social research documentation. In J. Kallas et al. (Eds.), *Development of infrastructures for data production and management in the social sciences*. [In Greek]. https://sodanet.gr/storage/publications/alexandreia_2022/1-Kallas_Paparisteidi.pdf

King, G. (2007). An introduction to the dataverse network as an infrastructure for data sharing. *Sociological Methods and Research, 36*(2), pp. 173-199. doi:10.1177/0049124107306660

Kondyli, D. & Klironomos, N. (2022). FAIR Data: Opportunities and challenges for research infrastructures and research communities. In J. Kallas et al. (Eds.), *Development of Infrastructures for data production and management in the social sciences*. [In Greek]. https://doi.org/10.17903/CV09INFRA

Kondyli, D., Nisiotis, C. S., & Klironomos, N. (2024). Data reusability for migration research: A use case from SoDaNet data repository. *Frontiers in Human Dynamics*, *5*. https://doi.org/10.3389/fhumd.2023.1310420

Linardis, A., Maravelakis, P., & Fragoulis, G. (2023). *Data collection methods using digital questionnaires and survey methodology. Management of online, in-person and telephone data and surveys with Limesurvey and SPSS*. Kallipos, Open Academic Publications. http://dx.doi.org/10.57713/kallipos-381

Linardis, A., & Ioannidis, A. (2022). Extending the life cycle of quantitative surveys and data. The role of data repositories. In J. Kallas et al. (Eds.), *Development of infrastructures for data production and management in the social sciences*. [In Greek]. https://doi.org/10.17903/CV02INFRA

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18