Generations and Gender Programme Preparatory Phase Project (GGP-5D)

REPORT ON FAIR DATA DOCUMENTATION AND DISSEMINATION IN THE GGP

Work package 1: **Technical Design**
Grant Agreement Number: **101079357**
Project acronym: **GGP-5D**
Project full title: **The Generations and Gender Programme Preparatory Phase Project**

Due delivery date: **1 April 2025**
Actual delivery date: **27 March 2025**
Organization name of lead participant for this deliverable**: French Institute for Demographic Studies (INED)**

Dissemination level: **Public**

Funded by the European Union

Document Control Sheet

Deliverable number:	D1.2
Deliverable responsible:	Laurent Toulemon
Work package:	1. Technical Design
Editor(s):	Vytenis Juozas Deimantas

Author (s)		
Name	***Organization***	***E-mail***
Thibaud Ritzenthaler	***INED***	***thibaud.ritzenthaler@ined.fr***
Laurent Toulemon	***INED***	***toulemon@ined.fr***
Arianna Caporali	***INED***	***arianna.caporali@recherche.gouv.fr***
Olga Grünwald	***NIDI***	***grunwald@nidi.nl***
Lars Dommermuth	***Statistics Norway***	***Lars.Dommermuth@ssb.no***

Document Revision History			
Version	***Date***	***Modifications Introduced***	
		Modification reason	***Modified by***
V1			
V2			
V3			

Executive summary

The Generations and Gender Programme organises consistent data collection on family behaviour in Europe, based on national comparative surveys, the Generations and Gender Surveys (GGS). This report presents the work on the GGP enhancing FAIR principles [1], aiming to make research data **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable, thereby facilitating collaboration, innovation, and knowledge sharing across disciplines and borders. To tackle the challenges of making its data FAIR, the GGP has developed a three-strategy approach.

The first strategy involves identifying a data repository that can implement the necessary measures to meet FAIR data principles. It conducted an exploratory study to identify a suitable data repository, pointing out Norwegian Agency for Shared Services in Education and Research (SIKT) as the preferred option based on a comparative analysis. As the proposed solution from SIKT proved too costly for the GGP at this stage, the GGP decided on developing a self-hosted GGP Data Portal by enhancing the existing Colectica setup, which offers benefits but also presents challenges. The long-term goal remains the transition to a SIKT-based solution, provided its affordability.

The second strategy involves assessing the FAIRness of the GGP data through existing methods, with a double aim of analysing the concept of FAIR data and evaluating the measures implemented in the framework of the first strategy. GGP's FAIRisation assessment [2] is informed by research highlighting the importance of Data Documentation Initiative (DDI) standards in achieving FAIRness and the need to increase FAIRness in all domains. Key actions are needed like using permanent Digital Object Identifiers (DOIs), controlled vocabularies, and external data/metadata links. The FAIR-Impact project [2] also provides insights, indicating that the GGP needs to apply these recommendations, while its software and metadata standards are well-established and recognised, requiring no changes.

The third strategy involves implementing measures to increase FAIRness at the level of data documentation, with a focus on controlled vocabularies in the GGP Colectica portal [3]. A two-phase plan was devised to implement a controlled vocabulary, which aims to enhance the FAIRness of metadata. This will improve the FAIRness of the data. The experience of the GGP in the GO-FAIR community, as well as the use of machine-actionable FAIR Implementation Profiles, will also be utilised to support this effort.

By the end of 2025, DOIs will be implemented and Creative Commons licensing will be adopted to enhance the discoverability, accessibility, and citation value of the GGP's data. Additionally, a strategic action plan will be finalised to further increase the FAIRness of the data and metadata, including the development of a knowledge graph to improve harvesting and integration of GGP data.

Table of Contents

Definitions and acronyms

CESSDA-ERIC: Consortium of European Social Science Data Archives
DANS: Data Archiving and Networked Services
DDI: Data Documentation Initiative
DOI: Digital Object Identifier
ELSST: European Language Social Science Thesaurus
ESS-ERIC: European Social Survey
FAIR: Findable, Accessible, Interoperable, and Reusable
FFS: Family and Fertility Survey
FIP: FAIR Implementation Profiles
GESIS: Leibniz Institute for the Social Sciences
GGP: Generations and Gender Programme
GGS: Generation and Gender Survey
INED: French Institute for Demographic Studies
NIDI: Netherlands Interdisciplinary Demographic Institute
OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting
ODISSEI: Dutch Open Data Infrastructure for Social Science and Economic Innovations
SIKT: Norwegian Agency for Shared Services in Education and Research

1 Introduction

The *Generations and Gender Programme* (GGP) is an international research initiative that provides extensive, open-access data to support social science research, coordinated by the *Netherlands Interdisciplinary Demographic Institute* (NIDI) in collaboration with various European research institutions. *Generation and Gender Survey* (GGS), its longitudinal panel study [4], [5], is conducted in 18 countries (as of March 2025) for each round and focuses on family dynamics, life course trajectories, and intergenerational relationships, offering both micro- and macro-level insights. This standardised survey collects detailed information on demographic behaviours, including employment, income, education, health, intentions, and attitudes, with participants being interviewed three times over a six-year period (in t, t+3 and t+6, t being the year of the first wave of interviews) [6]. By tracking social changes over time, the GGP provides valuable insights for academic research and evidence-based policymaking.

Alongside the GGS, a complementary international dataset, known as *Harmonized Histories*, has been developed. This dataset involves the standardisation and integration of existing survey data into a unified format, thereby facilitating comparative research across countries. The primary emphasis of *Harmonized Histories* lies in the examination of fertility and partnership histories, with the data being structured in a manner that lends itself to event history analysis. Moreover, the dataset encompasses a range of supplementary variables, including socio-economic status, geographical location, and information regarding the respondent's family of origin, such as parental divorce and sibling configuration.

The standardised data collection facilitates cross-national comparisons, allowing researchers to examine demographic changes across varying cultural and institutional contexts. Additionally, the survey provides a comprehensive perspective on gender roles and inequalities, contributing to discussions on social policy and gender equality. By addressing themes such as contraception, education, housing, and personal networks, the GGP serves as a crucial resource for scholars and policymakers seeking to understand and respond to contemporary social challenges. Its open-access data structure enables rigorous analysis of family and partnership trends, reinforcing its significance in demographic research as well as for researchers from other disciplines.

1.1 FAIR environment and the GGP

In recent years, there has been a growing emphasis on the importance of making research data *Findable, Accessible, Interoperable, and Reusable* (FAIR). The FAIR principles [1] aim to enhance the discoverability, accessibility, and usability of research data, thereby facilitating collaboration, innovation, and knowledge sharing across disciplines and borders. In this context, the European Union has launched several initiatives to promote the adoption of FAIR data practices and to develop the necessary infrastructure to support the sharing and reuse of research data.

As part of these efforts, it's been decided in the framework of the GGP-5D project to include a specific task aimed to enhance the FAIRness of existing data and related services. Task 1.1 of the GGP-5D project aims to improve the discoverability, accessibility, and reusability of data by applying permanent identifiers to all datasets, making documentation compliant with controlled vocabularies and metadata standards, selecting Creative Commons licenses, and implementing the *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH). This task, led by the *French Institute for Demographic Studies* (INED) in collaboration with the *Netherlands Interdisciplinary Demographic Institute* (NIDI) and the *Consortium of European Social Science Data Archives* (CESSDA-ERIC), has

already been validated and is now being implemented to ensure that research data is made widely available and usable for decision-makers and researchers across the European Union.

CESSDA-ERIC is a distributed research infrastructure in the social sciences domain. Its mission is to enable the research community to conduct high-quality research in the social sciences. This, in turn, contributes to the production of effective solutions to the major challenges facing society, both currently and in the future. CESSDA-ERIC is promoting FAIR principles by ensuring social science data is properly documented and indexed, making it easily findable through standardised metadata standards and data discovery tools. Also, the infrastructure facilitates access to data through secure repositories and access controls, while promoting interoperability through standardised data formats and APIs. By also encouraging data sharing, citation, and replication, CESSDA makes social science data more reusable, ultimately supporting high-quality research and innovation in the field.

The documentation of GGP survey data adheres to the international standard of the *Data Documentation Initiative* (DDI), specifically last DDI-Lifecycle version which is specifically designed for longitudinal studies. This compliance is in line with the recommendations of CESSDA-ERIC. The GGP online platform [7] (see some screenshots in Appendix 1) provides a comprehensive repository for browsing both the data and its corresponding documentation.

Country teams contribute to the GGP Central Coordination Team by providing fieldwork metadata and survey methodology information, which is subsequently reported on the online portal. The original questionnaire, available in PDF, DDI and web format, is accompanied by supplementary methodological documentation. Furthermore, each variable is thoroughly documented, encompassing labels, response categories, missing values, question texts, filters, and descriptions of country-specific deviations from the standard questionnaire. The calculation methods for derived variables are also provided. Additionally, users can browse variable tabulations, rendering the portal a multifaceted resource that serves as both an online codebook and a metadata repository for GGS surveys.

The use of DDI enables the survey metadata to be interoperable and machine-readable, facilitating efficient transfer to external archives that support DDI. This ensures seamless data integration and enhances the overall accessibility of the GGP survey data.
On the technical side, the data management infrastructure is built with Colectica, a centralised repository, based on DDI-Lifecycle, which provides a robust framework for managing data resources, facilitating collaborative workflows, and ensuring automatic version control. This repository serves as the foundation for data publication and discovery, enabling the dissemination of data and metadata through a web-based interface. Furthermore, the infrastructure supports the comprehensive documentation of data collection, survey specification, and dataset descriptions, allowing for seamless integration with existing repositories and promoting efficiency and collaboration throughout the data lifecycle.

1.2 Strategies to tackle the task

The initial strategy for attaining FAIR data within the GGP involves the identification of a data repository that is recognised by CESSDA-ERIC and is capable of ensuring compliance with the principles of FAIR data. Notably, the chosen repository should allow to assign a *Digital Object Identifier* (DOI) to the data, ensuring the unique identification of the datasets and facilitating their accurate citation. Moreover, the repository needs to be able to apply Creative Commons Licenses, which provide explicit guidelines governing data reuse and sharing, thus promoting transparency and clarity. Additionally, the repository should enable harvesting by the CESSDA-ERIC catalogue, which serves to facilitate data discovery and enhance accessibility.

The GGP also requires a repository that's flexible enough to adapt to data access procedures and needs, including its data access procedures. This approach implies that the GGP needs a repository that enhances data management, and ensures that the data are readily accessible and available for reuse, whilst also aligning with the programme's specific objectives and guidelines.

The second strategy involves assessing the FAIRness of GGP data through existing methods, with a twofold aim. Firstly, it seeks to analyse the concept of FAIR data by drawing on the work of researchers across disciplines who have evaluated FAIR and increased FAIRness, in order to efficiently implement the FAIR principles. Secondly, it aims to evaluate whether the measures identified in the first strategy were sufficient to make GGP data FAIR [2], with the collaboration of CESSDA-ERIC.

The third strategy involves implementing measures to increase FAIRness at the level of data documentation, with a focus on controlled vocabularies in the GGP Colectica portal. This approach enables the GGP to take immediate action to enhance the FAIRness of its data.
A key component of this strategy is the development of a controlled vocabulary work plan, which aims to build vocabularies that are most suited to the GGP's needs and then implement them. This involves creating a structured and standardised system for categorising and describing data, which will facilitate data discovery, accessibility, and reuse.

By implementing controlled vocabularies, the GGP can improve the consistency and accuracy of its data documentation, making it easier for users to find and understand the data. This, in turn, will enhance the overall FAIRness of the GGP's data, making it more accessible and reusable for researchers and other stakeholders.

2 Strategies put in place to tackle the challenges of the task

2.1 Looking for a data repository for the GGP

In order to address the challenges of Task 1.1, an exploratory study was conducted to identify a suitable data repository for the GGP. A comparative analysis of three data repositories, namely DANS (*Data Archiving and Networked Services*, the Netherlands), GESIS (*Leibniz Institute for the Social Sciences*, Germany), and SIKT (*Norwegian Agency for Shared Services in Education and Research*, Norway), was presented to the GGP Consortium Board in April 2023. The evaluation was based on a set of criteria that were deemed essential for the GGP's data management needs.

The criteria included the flexibility of data access procedures, the deposit system's ability to attribute DOIs, compliance with the CESSDA-ERIC metadata model, and DDI-Lifecycle compliance. Additionally, the repositories were assessed on their data preservation and curation capabilities. The cost of using each repository was also taken into consideration. Finally, the visibility of the data was evaluated, including the availability of an own landing page for the GGP in the data repository and the ability to be harvested by the CESSDA-ERIC catalogue.

Following a thorough evaluation of these criteria, SIKT emerged as the most suitable option, demonstrating superior performance across the assessed dimensions. The selection of SIKT as the preferred data repository was validated by the GGP Steering Committee in September 2023, based on its ability to meet the GGP's specific needs and requirements, ensuring the long-term preservation and accessibility of the programme's data.

As a provider of shared services, SIKT delivers high quality digital services, having the ambition to establish itself as a preferred provider and partner. It provides a comprehensive digital foundation for

the knowledge sector, enhances information security and data protection, and supports the realisation of strategies through user-centric solutions and shared services.
Following the validation of SIKT as the preferred data repository, the GGP Steering Committee and Consortium Board endorsed the GGP Central Coordination Team's proposal to switch to SIKT and develop a platform similar to that of the ESS-ERIC, including custom branding. However, this solution proved to be too expensive, and SIKT subsequently proposed a lower-cost temporary solution, *SurveyBanken*. Although *SurveyBanken* offered some advantages, such as outsourcing certain operations to SIKT, it did not meet the GGP's initial requirements.

As a result, the GGP began investigating alternative options. After careful consideration, it was decided to upgrade the current GGP setup and develop a GGP Data Portal, rather than opting for *SurveyBanken*. The new GGP Data Portal was created by improving the design and functionality of GGP Colectica Portal, enhancing the GGP User Space, and upgrading the backend framework (i.e., Django Admin). As the developments are made in line with FAIR and DDI principles, they will also enable a possible transition from the GGP data portal to SIKT at a later stage.

The benefits of creating a new self-hosted GGP Data Portal include increased branding opportunities, timely DOI assignment, and cost-effectiveness. Challenges associated with this solution include an increased workload and maintenance requirements, as well as the need for technical expertise to develop the portal.
Despite these challenges, the development of a self-hosted platform emerged as the best short-term solution, enabling the GGP to disseminate data continuously and effectively, increase the FAIRness of the data infrastructure and give GGP holders a better understanding of the requirements that a future data platform will have to meet. On the long term, a platform management by SIKT remains the preferred solution for the GGP, provided that it becomes affordable for the GGP (through specific funding or a collaboration with SIKT). The Norwegian GGP national team, in collaboration with SIKT and the GGP, had submitted a joint application in response to call from the Norwegian National Financing Initiative for Research Infrastructure, but despite positive evaluations, the application was not successful in securing funding in September 2024. A revised application will be submitted if a new call from the Research Council of Norway is released. In light of this outcome, the GGP has subsequently taken on the responsibility of developing the Data Portal, thereby ensuring the project's continuation and advancement.

This includes hosting the datasets at GGP Central Coordination Team, whilst DANS will assume responsibility for the long-term archiving of the datafiles. To facilitate seamless access, DOIs will be redirected to the new location upon long term preservation. This solution ensures the enduring preservation and accessibility of the GGP's data, whilst also fulfilling the requisite criteria for data archiving, thereby guaranteeing datasets' continued availability and integrity over time.

2.2 Assessment of FAIRisation in the GGP

The assessment of FAIRisation in the GGP informed by existing research [2] who highlighted the importance of DDI in achieving FAIRness, and the need to increase all domains of FAIRness. They also identified key actions needed, including the implementation of persistent identifiers, controlled vocabularies, and links to external data and metadata.

Research conducted as part of the project with FAIR-Impact [8], [9], [10] also provided valuable insights. The take-home message for the GGP is that it needs to apply persistent identifiers, controlled vocabularies, and improve external references. Colectica, the software and DDI, the metadata standard used by the GGP are well-established in the research community on FAIR-data and recognised by CESSDA-ERIC, so no changes are required and planned in this regard.

2.3 Increasing the GGP FAIRness

A two-phase plan was devised to facilitate the effective implementation of controlled vocabularies. The first phase involved an analysis of controlled vocabularies, which helped to elucidate its potential to enhance the FAIRness of metadata.

Controlled vocabularies refer to a set of predefined terms or words that are used to describe a specific concept, topic, or domain. They are a key approach to improve communication and data exchange among different groups. By establishing a standard set of terms, controlled vocabularies create a shared understanding, reducing the risk of ambiguity and inconsistency. This enables more effective data retrieval and analysis, leading to better insights and decision-making. [11] Controlled vocabularies are also becoming increasingly important in data management, particularly in fields with complex concepts and terminology, such as social sciences. In this domain, for survey such as the GGS, it can be sampling procedure, analysis units, country codes, data collection methodologies, etc …

The results indicate that the implementation of controlled vocabularies can significantly improve the findability of the data. This approach offers several practical advantages, including facilitating the retrieval of relevant information, standardising metadata across organisations, and ensuring completeness and consistency in online searches.

Controlled vocabularies can overcome language differences in online searches and are compatible with the DDI framework, which enables the production of interoperable metadata. On a technical note, the Colectica software, currently used for documentation, supports the implementation of controlled vocabularies, and a list of relevant controlled vocabularies has been identified for use in the project.

The second phase involved the selection of the controlled vocabularies adopted by CESSDA [12]. The aim was to select the largest panel of controlled vocabularies to provide a standardised framework for describing various aspects of the data, including the entity being analysed, geographic locations, data formats, collection methods, and thematic classifications. The full list of controlled vocabularies selected is available in Appendix 2. The use of these CESSDA-controlled vocabularies, including the *European Language Social Science Thesaurus* (ELSST), will facilitate the creation of high-quality, interoperable metadata that conforms to established standards and best practices in the field.

The experience of the GGP in the GO-FAIR community, in collaboration with the Dutch Open Data Infrastructure for Social Science and Economic Innovations (Odissei), has also been valuable. Odissei is a Dutch open data infrastructure for social science and economic innovations that centralises data collection, fostering methodological synergy and cost-efficiency through the coordinated application of shared standards and new technologies. By bringing together various international surveys and panels under one infrastructure, Odissei not only streamlines data collection but also reinforces robust documentation practices—such as those in the DDI-Lifecycle—that are essential for achieving FAIR.

The GO-FAIR community has developed machine-actionable FAIR Implementation Profiles (FIP), which are collections of FAIR implementation choices made by a community of practice for each of the FAIR Principles. The GGP has integrated this community with the help of Odissei, and has found that its documentation in DDI-Lifecycle is already a specific strength of the GGP. The implementation of DOIs is planned for 2025 and will contribute to increasing the FAIRness of the data.

3 Concrete measures put in place

3.1 Setup the controlled vocabularies

The implementation of CESSDA Controlled Vocabularies in the GGP Colectica Portal began with the creation of a JSON configuration (Figure 1) file for a single country, Norway, in a local environment. As a result of this implementation, DDI elements that are covered by controlled vocabularies are no longer open for free-text entry in Colectica Designer (local), but instead present a list of valid vocabulary terms, ensuring consistency and accuracy in metadata creation.
This involved a thorough review of the Colectica documentation to identify the correct "PropertyName", Colectica identification for controlled vocabulary integrated in the software, for each controlled vocabulary, which was then carefully specified in the configuration file. The "CodeList", referencing DDI object name, was also correctly linked to the DDI-L architecture, verifying that the unique identifier was associated with the corresponding DDI object.

```
{
 "Mappings": [
  {
   "PropertyName": "Methodology.SamplingProcedure",
   "CodeListId": "urn:ddi:int.ggp:c2d066d4-811e-4328-abb3-237b43722e41:1"
  },
  {
   "PropertyName": "StudyBase.AnalysisUnits",
   "CodeListId": "urn:ddi:int.ggp:dfd57235-edf7-48b6-ba48-72a423d67e86:1"
  }, …
 ]
}
```

Figure 1: *Extract of the JSON file of configuration*

The configuration file defined mappings for several key vocabularies, including Sampling Procedure, Analysis Units, Country Codes, Data Collection Methodologies, Time Methods, Type of Instruments, and Numeric Domain, all of which were drawn from the CESSDA Controlled Vocabularies (Figure 2). These mappings enabled the standardisation of key metadata elements, facilitating data discovery and reuse. Additionally, the Keywords property was mapped to the *European Language Social Science Thesaurus* (ELSST), a multilingual thesaurus for the social sciences that consists of over 3,400 concepts and covers the core social science disciplines (Figure 2).

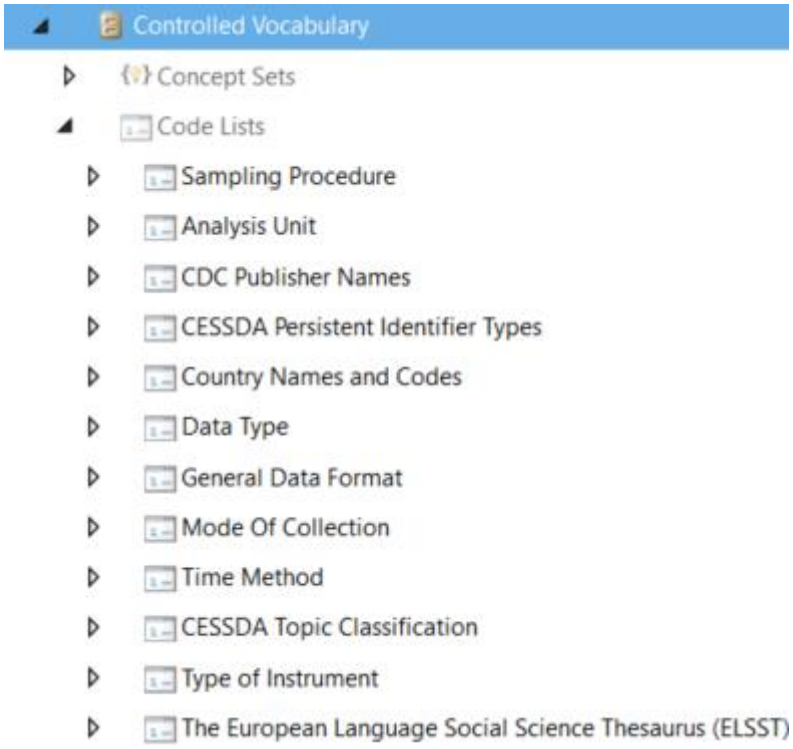


Figure 2: *DDI package of controlled vocabularies, loaded on Colectica*.

After publishing the configuration for Norway, technical issues were encountered and resolved through collaboration with the Colectica developers, resulting in improvements to the overall implementation. Thereafter, the process was replicated successfully for all GGP-countries, adapting the configuration file as needed to accommodate any variations in the controlled vocabularies.

By carefully configuring the controlled vocabulary, resolving technical issues, and adapting the implementation for multiple countries, CESSDA Controlled Vocabularies was successfully integrated into GGP Colectica Portal and ensured its seamless application across all countries (Figure 3).

Figure 3: *Keywords in DDI-L XML and in HTML on the documentation website*

3.2 Creating a new data portal

Following the reflections on how to best disseminate the data, the GGP decided to implement a new data portal with minimal investment. The decision was taken to create a self-hosted data portal. The primary objective of this new portal was to provide researchers with a single point of access to data and metadata, with the aim of enhancing their user experience.

To achieve this goal, the data portal is hosted on the main website of the GGP, leveraging the existing infrastructure and expertise. The development of the portal was undertaken by the same team who created the main GGP website, in close collaboration with the GGP Central Coordination Team. This approach enabled the GGP to create a functional and user-friendly data portal, while minimising investment and ensuring a seamless integration with the existing website. The portal is accessible at the following web address www.ggp-i.org/data-catalog/, and some pages are shown in Appendix 1.

The data portal was built on the Django platform, a high-level Python web framework that provides a robust and scalable foundation for web development. Django's modular design and extensive libraries enabled rapid development and deployment of complex web applications, making it an ideal choice for the GGP data portal. The use of Django also allows for easy integration with other tools and services, such as Colectica that is used for data documentation and metadata hosting. By leveraging the power and flexibility of Django, the GGP data portal is able to provide a robust and user-friendly interface for researchers to access and explore the data and metadata.

4 Plan to implement the measures not yet put in place

The strategic objectives for the year 2025 encompass several key initiatives: the implementation of DOIs, the adoption of Creative Commons [13] licensing, and the finalisation of a strategic action plan to enhance FAIR principles.
With the systematic implementation of DOIs, the GGP will assign persistent and unique identifiers to its datasets, thereby enhancing their discoverability and citation value. By utilising DataCite's Fabrica solution [14], the GGP Central Coordination Team will streamline the DOI attribution process, ensuring that the GGP data is accurately identified and accessible to researchers and stakeholders.

To foster openness and accessibility, the adoption of a default licence for the GGP metadata is under consideration. The *Creative Commons Attribution* (CC BY) licence is being explored as it permits the Open Access and chargeless use, sharing, and adaptation of metadata while mandating attribution to the original creators. Additionally, the *Creative Commons Attribution-ShareAlike* (CC BY-SA) licence is being evaluated, which stipulates that any derivatives of the metadata must be shared under identical licensing terms.

The strategic action plan to bolster the FAIRness of the GGP data will outline our approach to integrating DOIs, implementing Creative Commons licences, and developing an API to facilitate automated metadata harvesting from the data portal. Furthermore, the user interface of the data portal will be refined to enhance its intuitiveness and user-friendliness. To enhance these aspects, the first step would be for new members of the GGP Central Coordination Team to validate the practical use of the website. This will enable the identification of difficulties that users who are unfamiliar with the website may encounter. The next phase will be to set up a working group focused on improving the user interface of the Data Portal, based on the collective expertise of the GGP Central Coordination Team to address this task. Volunteer users will join the team to support this endeavour.

As part of the GRAPHIA project, a knowledge graph is planned for development by the end of 2025, in collaboration with GESIS (Leibniz Institute for the Social Sciences, Germany). A knowledge graph facilitates the standardisation and enrichment of metadata for Social Sciences and Humanities research objects, such as publications, archival materials, and cultural artefacts. It can overcome the challenges of heterogeneity and fragmentation. By employing persistent identifiers, standardised vocabularies, and ontologies, it enables the creation of rich, machine-readable metadata, thereby enhancing the findability and accessibility of data for both human researchers and automated agents [15], [16], [17].

This structured and standardised representation of data will enable the integration of the GGP datasets with other data sources and facilitate the discovery of relationships between data entities. Making *Harmonized Histories* and *Family and Fertility Survey* (FFS), the previous version of the GGP, metadata available on Colectica is scheduled to be implemented incrementally as needed. In long-term, the transfer of the Data Portal to SIKT will eventually be done. It will be eased by the use of standards such as DDI, controlled vocabularies and DOIs or Colectica software.

5 References

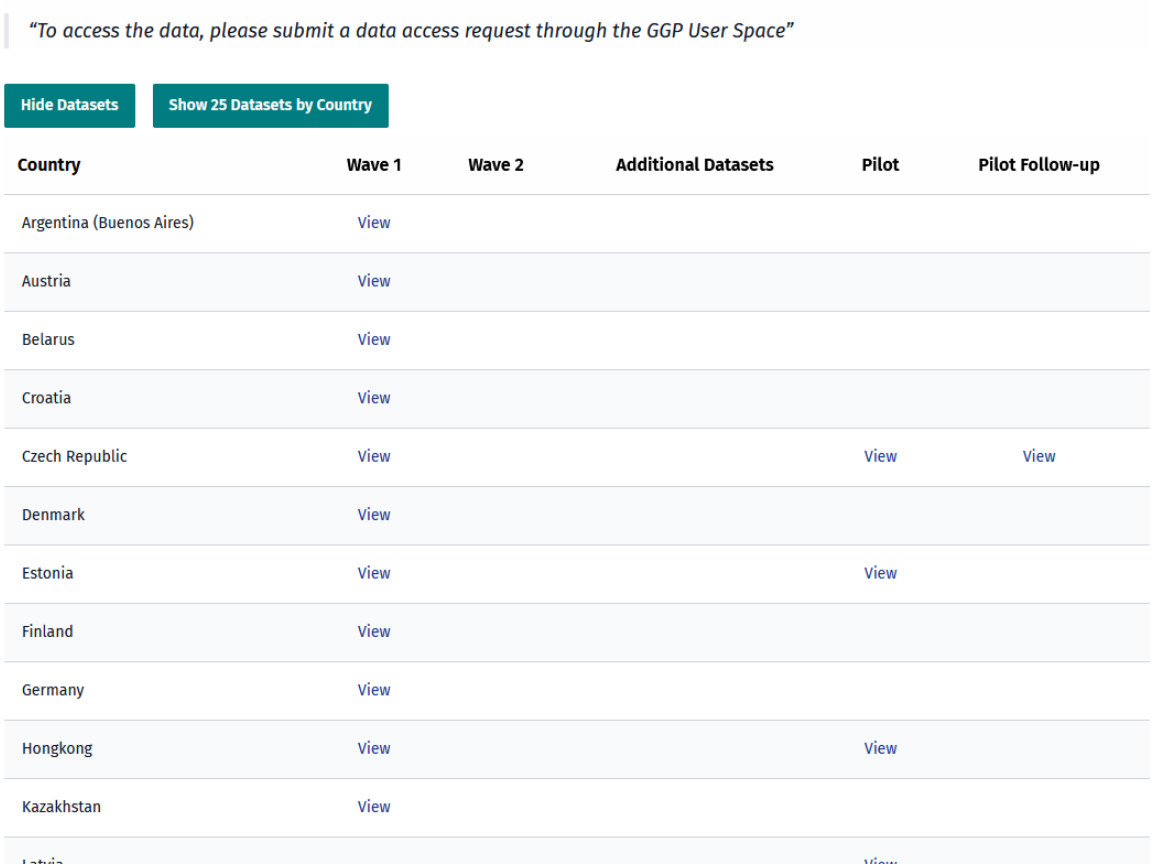
[1] M. D. Wilkinson *et al.*, 'The FAIR Guiding Principles for scientific data management and stewardship', *Sci. Data*, vol. 3, no. 1, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.
[2] 'GGP | Generations & Gender Programme FIP'. Accessed: Feb. 21, 2025. [Online]. Available: https://fip-wizard.ds-wizard.org/wizard/projects/807eae45-f353-4db9-ac7e-254debc4da0f

[3] C. Linés, A. Caporali, and O. Grünwald, 'FAIRness assessment of Generations and Gender Programme longitudinal survey dataset', Mar. 18, 2024. doi: 10.5281/zenodo.11402267.
[4] J. Iverson, 'Controlled Vocabularies in Colectica', Nov. 30, 2021. doi: 10.5281/zenodo.5749087.
[5] A. Vikat *et al.*, 'Generations and Gender Survey (GGS): Towards a better understanding of relationships and processes in the life course', *Demogr. Res.*, vol. 17, pp. 389–440, Nov. 2007, doi: 10.4054/DemRes.2007.17.14.
[6] A. H. Gauthier *et al.*, 'Data Brief: The Generations and Gender Survey second round (GGS-II)', GGP Technical Paper Series, Nov. 2023. doi: 10.5281/zenodo.10220746.
[7] A. Gauthier *et al.*, 'GGP Technical Guidelines', Mar. 2024, Accessed: Feb. 24, 2025. [Online]. Available: https://zenodo.org/records/10812889
[8] 'Data Portal – GGP'. Accessed: Feb. 24, 2025. [Online]. Available: https://www.ggp-i.org/data-portal/
[9] A. Jacobsen *et al.*, 'FAIR Principles: Interpretations and Implementation Considerations', *Data Intell.*, vol. 2, no. 1–2, pp. 10–29, Jan. 2020, doi: 10.1162/dint_r_00024.
[10] E. Schultes, B. Magagna, K. M. Hettne, R. Pergl, M. Suchánek, and T. Kuhn, 'Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence', in *Advances in Conceptual Modeling*, vol. 12584, G. Grossmann and S. Ram, Eds., in Lecture Notes in Computer Science, vol. 12584. , Cham: Springer International Publishing, 2020, pp. 138–147. doi: 10.1007/978-3-030-65847-2_13.
[11] A. M. Maineri, 'Controlled vocabularies for the social sciences: what they are, and why we need them', Oct. 2022, doi: 10.5281/zenodo.7157800.
[12] 'CESSDA Vocabulary Service'. Accessed: Feb. 28, 2025. [Online]. Available: https://vocabularies.cessda.eu/
[13] 'About CC Licenses', Creative Commons. Accessed: Feb. 28, 2025. [Online]. Available: https://creativecommons.org/share-your-work/cclicenses/
[14] 'DataCite Fabrica', DataCite Fabrica. Accessed: Feb. 28, 2025. [Online]. Available: https://doi.datacite.org/
[15] F. Beretta, 'Semantic Data for Humanities and Social Sciences (SDHSS): an Ecosystem of CIDOC CRM Extensions for Research Data Production and Reuse', 2024. doi: 10.33968/9783966270502-05.
[16] S. Mohamed *et al.*, 'Knowledge Graphs: The Future of Data Integration and Insightful Discovery', Dec. 17, 2024, *arXiv*: arXiv:2502.15689. doi: 10.48550/arXiv.2502.15689.
[17] 'Knowledge Graphs, AI Services and Next Generation Instrumentation for Research and Development in Social Sciences and Humanities | GRAPHIA Project | Fact Sheet | HORIZON', CORDIS | European Commission. Accessed: Feb. 28, 2025. [Online]. Available: https://cordis.europa.eu/project/id/101188018

6 Appendix

6.1 Appendix 1. Screenshots of the new GGP Data Portal

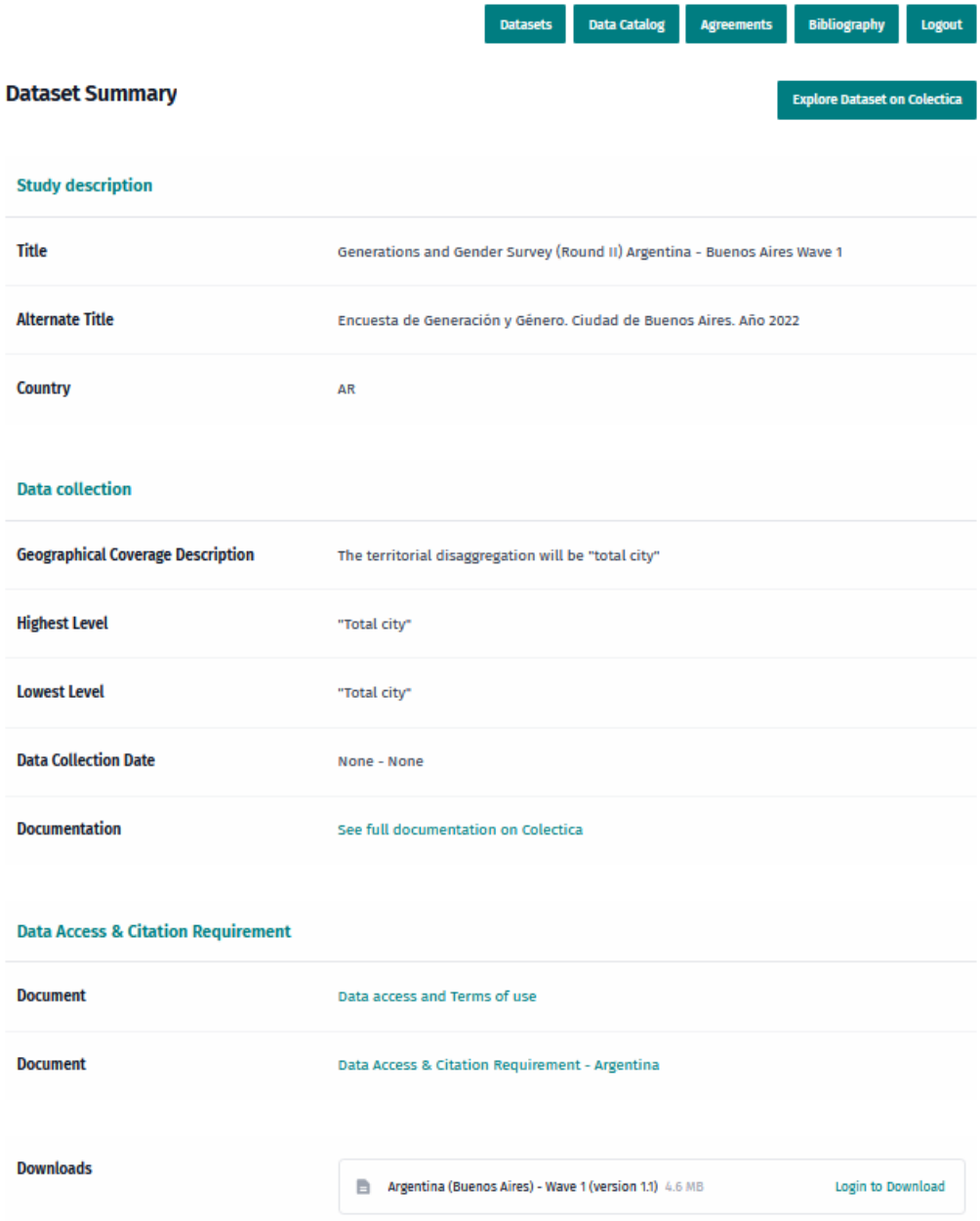
The GGP Data Portal (www.ggp-i.org/data-portal) is an extensive repository of datasets curated by the GGP, encompassing resources such as the GGS, Harmonised Histories, and the Family and Fertility Survey. This catalogue provides researchers with access to comprehensive data on family dynamics, life course trajectories, and intergenerational relationships across various countries. To facilitate scholarly analysis, the catalogue includes links to detailed documentation and offers datasets freely for non-commercial purposes upon registration in the GGP User Space.



"To access the data, please submit a data access request through the GGP User Space"

Hide Datasets | Show 25 Datasets by Country

Country	Wave 1	Wave 2	Additional Datasets	Pilot	Pilot Follow-up
Argentina (Buenos Aires)	View				
Austria	View				
Belarus	View				
Croatia	View				
Czech Republic	View			View	View
Denmark	View				
Estonia	View			View	
Finland	View				
Germany	View				
Hongkong	View			View	
Kazakhstan	View				
Latvia				View	



This screenshot of the data catalogue shows that the Generations and Gender Survey (Round II) Argentina - Buenos Aires Wave 1 dataset provides complete data collected from Buenos Aires in 2022. This dataset offers valuable insights into family dynamics, life course trajectories, and intergenerational relationships within the context of Buenos Aires. Researchers can access detailed documentation and explore the dataset through the Colectica portal.

6.2 Appendix 2. List of the selected controlled vocabulary

CV Name	CESSDA Definition
Analysis Unit	Describes the entity being analysed in the study or variable. This vocabulary can also be used to describe the unit of observation, which is the unit being observed, or from which data are collected. The unit of observation can be the same as, or different from the unit of analysis.
Country Names And Codes	ISO 3166-1 alpha-2 country codes and names.
Mode Of Collection	The procedure, technique, or mode of inquiry used to attain the data.
Sampling Procedure	A typology of sampling methods.
Time Method	Describes the time dimension of the data collection.
Type Of Instrument	Includes a typology of data collection instruments.
Topic Classfication	A typology of main themes or subjects of data.