Research and Innovation Action

Social Sciences & Humanities Open Cloud

Project Number: 823782 Start Date of Project: 01/01/2019 Duration: 40 months

Deliverable 9.5

Feasibility report on setting up a collection on questionnaires relating to Ethnic and Migrant Minorities in the European Question Bank

Dissemination Level	PU
Due Date of Deliverable	31/12/2021 (M36)
Actual Submission Date	15/12/2021
Work Package	WP9 - Data Communities
Task	T9.2 Ethnic and Migration Studies
Type	Report
Approval Status	Waiting EC approval
Version	V1.0
Number of Pages	p.1 – p.53

Abstract:

This report assesses the feasibility of setting up, as part of the CESSDA ERIC European Question Bank (EQB), a collection dedicated to the Ethnic and Migrant Minorities (EMM) surveys identified via the development of the EMM Survey Registry. Based on the outcomes of this assessment, the report argues that an EMM collection for the EQB (to be set up via harvesting by the EQB from the EMM Question Data Bank (QDB)), can be created if the various resource demands are met, and the link between the EMM QDB and EQB can be established in a sustainable and durable manner.

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the EU Horizon 2020 Research and Innovation Programme (2014-2020); H2020-INFRAEOSC-04-2018, under the agreement No. 823782

History

Version	Date	Reason	Revised by
0.0	20/10/2021	Initial outline of the report	Ami Saji; Laura Morales; Meredith Winn
0.1	24/11/2021	First draft of the report	Ami Saji; Laura Morales; Meredith Winn
0.2	07/12/2021	Peer review	Iris Alfredsson; Veronika Heider; Carsten Thiel
0.3	13/12/2021	Final draft using peer review feedback	Ami Saji; Laura Morales
1.0	14/12/2021	Final version for submission	Ami Saji; Laura Morales

Author List

Organisation	Name	Contact Information
Sciences Po	Laura Morales	laura.morales@sciencespo.fr
Sciences Po	Ami Saji	amikatherine.saji@sciencespo.fr
Sciences Po	Meredith Winn	meredith.winn@sciencespo.fr

Executive Summary

This document is a deliverable of the Social Sciences and Humanities Open Cloud (SSHOC) project. It is a report assessing the feasibility of setting up, as part of the CESSDA ERIC-led European Question Bank (EQB), a collection dedicated to the questionnaires and question items included in quantitative surveys on Ethnic and Migrant Minorities' (EMMs') inclusion and/or integration identified as part of the creation of the EMM Survey Registry (i.e. the *SSHOC Deliverable 9.4 Database with the metadata of surveys to EMMs across Europe*[1]). As such, the main aim of this report is to provide a detailed account of how the ethnic and migration studies data community—i.e. the team of Task 9.2 Ethnic and Migrant Studies of Work Package 9 Data Communities of SSHOC, COST Action 16111 - ETHMIGSURVEYDATA, and the French Open Science project, FAIRETHMIGQUANT—critically examined the realities of being able to create an EMM collection for the EQB.

The report begins by illustrating how the FAIR principles, as well as the cumulative experience in developing the EMM Survey Registry, led to the EMM collection for the EQB being conceptualized as a standalone service, the EMM Question Data Bank (QDB) from which the EQB would be able to ingest the questionnaires of the EMM surveys. It then describes the key challenges identified in setting up the conceptualized EMM collection and the subsequent solutions to these challenges, which in turn helped create workflows for creating an EMM collection. It concludes with a reflection of the overall assessment process, undertaken by the ethnic and migration studies data community, where the argument is made that an EMM collection for the EQB can be constructed, but with careful consideration of the budgetary, human resources, and technical demands of creating such an innovative and novel resource, as well as the sustainability and durability of the linkage between the EMM QDB and the EQB.

[1] Ami Saji, & Laura Morales. (2020). D9.4 Database with the metadata of surveys to EMMs across Europe (v1.0). Zenodo. https://doi.org/10.5281/zenodo.4558307

Abbreviations and Acronyms

ANR	Agence nationale de la recherche
API	Application Programming Interface
CDSP	Center for Socio-Political Data
CEE	Centre for European Studies and Comparative Politics
CESSDA ERIC	Consortium of European Social Science Data Archives European Research Infrastructure Consortium
CESSDA MDO	Consortium of European Social Science Data Archives Metadata Office
CESSDA MO	Consortium of European Social Science Data Archives Main Office
CMM	CESSDA Metadata Model
CNRS	Centre national de la recherche scientifique
DDA	Dansk Data Archive
DDI	Data Documentation Initiative
DeZIM	Das Deutsche Zentrum für Integrations- und Migrationsforschung
DNA	Danish National Archives
EC	European Commission
EKKE	National Center of Social Research
EOSC	European Open Science Cloud
EMM	Ethnic and migrant minority
EQB	European Question Bank
ESS-UPF	European Social Survey, Universitat Pompeu Fabra Barcelona
ETHMIGSURVEYDATA	International Ethnic and Immigrant Minorities' Survey Data Network, COST Action 16111
FAIR	Findable, Accessible, Interoperable, Re-usable
FAIRETHMIGQUANT	Making Ethnic and Migrant Minority Survey Data FAIR project (ANR Science Ouverte)
FTE	Full-time employee
GDPR	General Data Protection Regulation
GESIS	Leibniz Institute for the Social Sciences

GGP	Generations & Gender Programme
ICPSR	Institute for Social Research at the University of Michigan
INED	Institut national d'études démographiques
ISO	International Organization for Standardization
LOCALMULTIDEM	Multicultural Democracy and Immigrants' Social Capital in Europe project, FP6
NLP	Natural language processing
NSD	Norwegian Centre for Research Data
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
QDB	Question data bank
QVDB	Questionnaire Variable Database
RA	Research assistant
SP	Service provider
SSHOC	Social Sciences and Humanities Open Cloud
SSVD	Social Science Variables Database
VPN	Virtual private network
XML	Extensible markup language

Table of Contents

1. Introduction and background

SSHOC Task 9.2 Ethnic and Migrant Studies of Work Package 9 Data Communities contributes work to include tools and resources generated by the ethnic and migration studies data community (i.e, the Task 9.2 team based at the Centre for European Studies and Comparative Politics (CEE) of Sciences Po (France), COST Action 16111 - ETHMIGSURVEYDATA, and the French Open Science project, FAIRETHMIGQUANT), as well as to ensure that these tools and resources are, whenever possible, well represented in the SSHOC Catalogue and Marketplace and in EOSC.[2] Its main objective is thus to make quantitative survey data on Ethnic and Migrant Minorities' (EMMs') integration and/or inclusion,[3] in Europe and beyond, FAIR (Findable, Accessible, Interoperable, Re-usable).[4] To achieve this objective, the ethnic and migration studies data community has committed to delivering two distinct outputs as part of SSHOC:

1. Launching a new service and tool: the EMM Survey Registry,[5] a free online searchable database that is FAIR-compliant and displays compiled survey-level metadata for existing EMM surveys from over 30 different European countries;[6] and
2. Assessing the feasibility of setting up, as part of the CESSDA ERIC-led European Question Bank (EQB), a collection dedicated to the EMM surveys identified via the development of the EMM Survey Registry.

In order for these two outputs to be successfully pursued, SSHOC has provided the Task 9.2 team with essential funding and technical support. Additionally, the Task 9.2 team has mobilized durable collaborations with two different initiatives also striving to make quantitative EMM survey data FAIR. The first is COST Action 16111 – ETHMIGSURVEYDATA (International Ethnic and Immigrant Minorities' Survey Data Network), a European research network funded by the COST Association and made up of over 200 EMM-focused (non)academic researchers;[7] ETHMIGSURVEYDATA's primary contribution has been to provide the intellectual drive and momentum needed for realizing both of the outputs and constitutes the underlying research community that supports in the background the goals implemented by the Task 9.2 team. The second is FAIRETHMIGQUANT, an Open Science project funded

[2] Access to the Ethnic and Migration Studies data community page on the SSHOC website: https://www.sshopencloud.eu/tagging/ethnic-and-migrant-minorities [10 November 2021]
[3] From this point forward, the term, *EMM surveys*, is used to refer to quantitative surveys examining the integration and/or inclusion of EMMs.
[4] Description of the FAIR principles: https://www.go-fair.org/fair-principles/ [10 November 2021]
[5] Access to the landing page of the live version of the EMM Survey Registry: https://ethmigsurveydatahub.eu/emmregistry/ [10 November 2021]
[6] The 30 plus countries originally committed to contributing to the EMM Survey Registry are those who joined and participated in COST Action 16111 ETHMIGSURVEYDATA: https://www.cost.eu/actions/CA16111/#tabs+Name:Parties [10 November 2021]. Throughout this report, all countries that joined and participated in ETHMIGSURVEYDATA are referred to as 'European,' as ETHMIGSURVEYDATA is a European-led or focused research network; however, it is important to note that Turkey and Israel are part of the 30 plus countries. Moreover, due to resource constraints and challenges, Israel was unable to contribute metadata to the EMM Survey Registry, before ETHMIGSURVEYDATA concluded on 20 October 2021.
[7] COST Actions are research networks funded by COST (European Cooperation in Science and Technology) and countries (primarily those from Europe) formally join COST Actions to support their respective research objectives and goals. COST Actions are intended to foster research collaboration and innovation through networking activities (e.g. meetings, conferences, training schools, short-term research stays, publications, dissemination activities); as such, COST Actions are unable to fund the actual costs of undertaking research (e.g. human resources or technical/technological development). More information about ETHMIGSURVEYDATA can be found on this webpage of the COST website: https://www.cost.eu/actions/CA16111/

by the French Agence nationale de la recherche (ANR); this project's main responsibilities have been to ensure the inclusion of the French EMM surveys in the two outputs, liaise with the French data research communities, as well as to co-lead the effort to test and trial the workflows for the second output.

With the resources provided by SSHOC and the partnerships established with ETHMIGSURVEYDATA and FAIRETHMIGQUANT, the ethnic and migration studies data community was able to successfully achieve the first output in October 2020. As detailed in the corresponding report,[8] the launched beta version of the EMM Survey Registry had fully functional front- and back-ends,[9] where all of the desired or core functionalities were available for use and had made publicly available metadata for 435 surveys from 11 different European countries.

Subsequently, by having successfully prepared the first output, the ethnic and migration studies data community was able to make two significant strides in their work. First, it has been able to substantially expand the EMM Survey Registry's metadata collection to cover roughly 1,700 surveys from 33 different European countries,[10] as well as ensure that the final version[11] of this database and tool can be delivered in winter 2021. Second, and more importantly, it has been able to shift its focus towards the second output (starting in fall 2020), as it had completed the intensive development work for the EMM Survey Registry, and it had a clearer understanding of the landscape of EMM surveys potentially relevant for the second output. Indeed, establishing the pilot for the second output required being able to select surveys from the EMM Survey Registry from those deemed of enough quality, focusing on specific pilot-relevant topics, and for which the questionnaires would be readily and publicly available.

This report is thus dedicated to the realization of the second output: a comprehensive assessment of the feasibility of setting up an EMM collection as part of the EQB, a CESSDA ERIC-led tool intended to allow users to discover and learn about questionnaires used in surveys held by a CESSDA Service Provider (SP). To reflect the rigorous and multiphase feasibility assessment conducted by the ethnic and migration studies data community, this report is structured into the following sections:

- **Conceptualizing a collection for the EQB dedicated to EMM surveys:** A summary of how the FAIR principles, the number and diversity of the EMM surveys detected as part of the EMM Survey Registry work, and the curation and ingestion workflow of the EQB led to the decision of creating an EMM Question Data Bank (QDB), from which the EQB could harvest metadata for the questionnaires of EMM surveys.
- **Identifying and overcoming the key challenges of creating a collection for the EQB dedicated to EMM surveys:** An overview of the key challenges identified, based on how the

[8] Ami Saji, & Laura Morales. (2020). D9.4 Database with the metadata of surveys to EMMs across Europe (v1.0). Zenodo. https://doi.org/10.5281/zenodo.4558307

[9] The front-end of the EMM Survey Registry refers to the interface where all the publicly available metadata can be accessed: https://registry.ethmigsurveydatahub.eu/ [10 November 2021]. Whereas, the back-end of the EMM Survey Registry corresponds to the interface where the metadata can be managed: https://registry.ethmigsurveydatahub.eu/nova/login [10 November 2021]. Unlike the front-end, which is fully accessible to the public, the back-end has restricted access. Full access to the back-end is limited to the administrators of the EMM Survey Registry (i.e. the Task 9.2 team). All other users (e.g. data producers) must be vetted and approved for access and can only use features of the back-end that allow them to contribute new metadata.

[10] The current metadata offerings of the EMM Survey Registry substantially exceeds the initial estimate: roughly 800 surveys from the 35 different European countries.

[11] The final version is currently defined as a version of the EMM Survey Registry where all the metadata committed by the 34 European countries is made publicly available. The present estimation is that the final version will display metadata for roughly 1,700 surveys from the 34 different European countries.

EMM QDB had been conceptualized, followed by a discussion of how the ethnic and migration studies data community strategically chose to address and handle each of the identified key challenges.

- **Determining the workflows for the EMM QDB:** A description of how the ethnic and migration studies data community established thematic pilots and used one of them to develop the workflows for retrieving and storing questionnaires, as well as documenting the questionnaires and metadata using Colectica Questionnaires and Colectica Designer.
- **Linking the EMM QDB to the EQB:** An explanation of how and why the EMM QDB will be set up on a Colectica Portal managed by the CDSP (Center for Socio-Political Data) of Sciences Po—a member of PROGEDO[12] (the French SP of CESSDA) and a participant of SSHOC, followed by an overview of how the EQB will subsequently harvest the metadata for questionnaires of EMM surveys via the EMM QDB set up on the CDSP's Colectical Portal.
- **Evaluating the feasibility of setting up a collection for the EQB dedicated to EMM surveys:** A discussion about the realities of setting up an EMM collection for the EQB, based on the experiences and insights gained by ethnic and migration studies data community through the testing and trialing done via the thematic pilot.

It is important to point out that many of the aforementioned phases overlapped and were executed in parallel. The processes for linking the EMM QDB to the CDSP's Colectica Portal and the EQB are still undergoing testing and that an update to the report will be published to reflect the final outcomes of the tests. Nevertheless, this report is still able to provide a comprehensive and reliable assessment of how feasible it is to set up a collection for the EQB that would be dedicated to the EMM surveys.

2. Conceptualizing a collection for the EQB dedicated to EMM surveys

ETHMIGSURVEYDATA, which predates SSHOC, had an original interest in creating a service that would allow users to discover and learn from questionnaires from existing EMM surveys. The main purpose of such a service was (and still is) to provide a single access point through which the large interdisciplinary community of ethnic and migration studies researchers could search for existing questionnaire items that have been used in past studies to capture any given dimension of EMMs' integration and/or inclusion.

With this purpose serving as a guiding force, ETHMIGSURVEYDATA subsequently refined the goals of this service. First, it wanted to provide a resource that would, on the one hand, allow replicability of previous questionnaire items in future questionnaires and surveys and, on the other hand, to facilitate post-hoc comparisons of surveys that had been independently undertaken. Second, it wanted to promote more systematic and comparative knowledge consolidation in the field of ethnic and migrant studies within the social sciences.

Once the SSHOC project began, the service was further articulated as setting up a collection for the EQB dedicated to EMM surveys. Given that CESSDA ERIC would invest in the EQB as a sustainable tool, it was

[12] Access to the PROGEDO website: http://www.progedo.fr/en/ [10 November 2021]

strategic to efficiently allocate and invest resources to contribute to this wider tool that was likely to become very well known across all research communities in the SSH.

This section therefore discusses how the ethnic and migration studies data community conceptualized a collection for the EQB dedicated to EMM surveys. It first examines the important role played by the FAIR principles in informing and determining the main characteristics of the EMM collection. It then describes how both the number and diversity of EMM surveys, uncovered as part of the EMM Survey Registry work (i.e. the first output), and the curation and ingestion workflow of the EQB ultimately led to the decision of creating a new tool, the EMM QDB, as well as defining the EMM collection as the set of EMM surveys to be harvested by the EQB from the EMM QDB.

2.1 Leveraging the FAIR Principles

As with the EMM Survey Registry, the EMM collection of the EQB was conceived to be a publicly available service that would help make quantitative survey data on EMMs FAIR. It was thus a natural choice to use the FAIR principles themselves in defining the main characteristics of the EMM collection. Specifically, through information gleaned from consulting existing question data banks (e.g. Social Science Variables Database (SSVD) by ICPSR,[13] Dansk Data Archive (DDA),[14] CLOSER Discovery by CLOSER[15]), deliberations held by the ethnic and migration studies data community, and preliminary discussions conducted with the SSHOC partners leading the EQB development work—i.e. CESSDA ERIC and GESIS, the ethnic and migration studies data community determined that the EMM collection would represent the FAIR principles as indicated in **table 1** below.

Table 1: Representation of the FAIR Principles by the EMM collection of the EQB

Principle	Corresponding main characteristics
Findable	• For each EMM survey included in the EMM collection, allow a user to easily find and view the corresponding questionnaire(s) and its characteristics (i.e. different location/country and language variations). • Allow a user to view and navigate all the items (i.e. questions, code lists or response options/categories, statements/instructions) included within a questionnaire. • Allow a user to easily search for and locate questionnaires and questionnaire items (particularly questions) of interest by offering basic search functionalities (e.g. keyword search). • Allow a user to find questions belonging to a specific concept by establishing a detailed and structured conceptual hierarchy of the core dimensions of integration (as well as other relevant conceptual hierarchies) that can be used to indicate conceptual links between questions within a questionnaire and from different questionnaires.
Accessible	• For all users, allow all the questionnaires and questionnaire items to be freely accessible. • Ensure that the questionnaires and questionnaire items are documented

[13] Access to the SSVD landing page: https://www.icpsr.umich.edu/web/pages/ICPSR/ssvd/ [11 November 2021]
[14] Access to the DDA (now the Danish National Archives (DNA)) question data bank's landing page: http://dda.dk/simple-search?lang=en [11 November 2021]
[15] Access to the CLOSER Discovery landing page: https://discovery.closer.ac.uk/ [11 November 2021]

	and presented in an easy-to-read and easy-to-understand manner.
Interoperable	• Document the questionnaires and questionnaire items so that the corresponding metadata can be harvested by other social sciences services (e.g. via an API), particularly those dealing with survey research. • Ensure that DDI standards are used to document the questionnaires, given the widespread preference for DDI in the social sciences.
Re-usable	• In addition to providing the exact wording of a question, provide users with information about order and placement (as well as instructions to interviewers and respondents, when applicable), so they can better assess when it would be appropriate to re-use or draw inspiration from a specific question. • For questionnaires that are not available in English (i.e. were not translated into English or were not a translation of an English-language master version), offer an English translation for each questionnaire item to maximize their potential for re-use. • Whenever possible, provide users with the URL/link, persistent identifier (e.g. DOI), and citation information for the original questionnaire.

Once these main and FAIR-informed characteristics of the EMM collection were established, the ethnic and migration studies data community liaised with experts, most notably from the SSHOC community, who could help further articulate the envisioned EMM collection. These knowledge sharing exchanges, illustrated, for the first time, the advantage of setting up the envisioned EMM collection by creating an EMM dedicated question data bank from which the EQB could harvest the metadata for questionnaires of EMM surveys (see section 2.2, Discovering a need for a EMM Question Data Bank (QDB), for more detailed information). They also supported the ethnic and migration studies data community in pinpointing specific challenges to be addressed in realizing the envisioned EMM collection (see section 3, Identifying and overcoming the key challenges of creating a collection for the EQB dedicated to EMM surveys, for the specifics). Finally, they clarified that in order to promote interoperability, the EMM questionnaires and all of its questionnaire items would need to have their metadata documented in DDI Lifecycle (as opposed to DDI Codebook, which was what was adopted for the EMM Survey Registry), as it is the metadata documentation standard being used by the EQB and is the one designed for social sciences survey research to capture relational information (i.e. conceptual links between questions).

While these discussions did not change how the EMM collection would ideally showcase the FAIR principles, it did clarify how the EMM collection would actually be created in practice, and how the next steps in terms of implementation were narrowed down in terms of options. Specifically, as also described in section 2.2, the decision was made to develop an EMM Question Data Bank (QDB): an independent product linked to a CESSDA SP, from which the EQB could harvest metadata for the questionnaires of EMM surveys. Thus, once the alpha version of the EQB[16] was released and planning was actively underway to launch the beta version, the ethnic and migration studies data community met with CESSDA Main Office (MO) and GESIS to assess that the EMM collection, which would now be set up via the EMM QDB, was still aligned with the in-development EQB. As a result, the ethnic and migration studies data community was able to confirm that EQB was being designed in a way that would be generally compatible with the envisioned EMM collection; the two main points of ambiguity

[16] The alpha version of the EQB was made accessible via the following URL: https://euroquestionbank-dev.cessda.eu/. At the time of writing, this URL is no longer active.

were the EQB's ability to handle multiple languages (as they had only been testing English and German questionnaires), as well as display rich conceptual information.

Therefore, the ethnic and migration studies data community was ultimately able to determine, from strictly a conceptual point of view, that an EMM collection responsive to the needs of the ethnic and migration studies data community, could be set up as part of the EQB. This, in turn, allowed them to focus more on testing and trialing the solutions identified for overcoming each of the challenges specified in section 3, Identifying and overcoming the key challenges of creating a collection for the EQB dedicated to EMM surveys.

2.2 Discovering a need for an EMM Question Data Bank (QDB)

Once the majority of the countries had identified the EMM surveys (in their respective country) to be covered by the EMM Survey Registry, the ethnic and migration studies data community examined the roughly 450 detected surveys to estimate the share of EMM surveys relevant to the EMM collection (i.e. those with a questionnaire that are available for re-use), as well as understand the variety in questionnaires for the relevant EMM surveys. What they subsequently learned was the following:

- Roughly 50% of the EMM surveys had the questionnaire available for re-use, either because it was made available through a public link or via request from a member of ETHMIGSURVEYDATA.
- Countries with a social science data archive and/or well-established practices for data archiving and preservation (e.g. Germany, Norway) were more likely to have EMM surveys with the questionnaire available for re-use.
- Roughly 50% of the EMM surveys with the questionnaire available for re-use offered it in at least one EMM-spoken language (i.e. languages spoken by EMM populations, such as Arabic, Chinese, Pashto, etc.).

With these insights, coupled with the then estimation that the final version of the EMM Survey Registry could cover around 800 surveys from the 35 European countries, it became evident that there could be a strong case for creating a question data bank dedicated to EMM surveys, in addition to setting up a collection for the EQB for the EMM surveys. The ethnic and migration studies data community was thus able to test this hypothesis by conducting intakes during internal meetings. The collective feedback they amassed indicated that researchers and policy-oriented professionals would benefit from having an EMM-specific question data bank for two principal reasons: (1) users of EMM surveys would be able to more efficiently find questionnaires and questions of interest, as the tool would only be covering EMM surveys; and (2) a standalone question data bank would, if needed, allow for the inclusion of features specific to the needs of users of EMM surveys (i.e. those not necessarily relevant for the average user of the EQB).

As a next step, the ethnic and migration studies data community met with CESSDA MO and GESIS, who advised that if the EMM-dedicated question data bank could be set up as an independent product linked to a CESSDA SP, the EMM collection could come to fruition by the EQB simply harvesting the metadata from the EMM-dedicated question data bank. Their argument was based on the fact that the

EQB would, as a first priority, be populated by harvesting metadata in DDI Lifecycle for questionnaires of surveys held by a CESSDA SP.

Ultimately, the ethnic and migration studies data community decided to create the EMM Question Data Bank (EQB), in order to respond to user feedback about the merits of having a dedicated question data bank for EMM surveys, as well as to facilitate the process of setting up an EMM collection. Specifically, the EMM QDB would be a question data bank for EMM surveys only, operate autonomously from the EQB, and be set up in partnership with a CESSDA SP (i.e. the CDSP, as explained in more detail in section 3.4, Challenge 4: Creating a public-facing interface for the EMM QDB).

Finally, to properly pursue the development of the EMM QDB, the ethnic and migration studies data community expanded from the Task 9.2 team and ETHMIGSURVEYDATA to include the Open Science project funded by the ANR: FAIRETHMIGQUANT. The inclusion of this new project formally ensured the support of the CDSP (a member of PROGEDO, the French CESSDA SP) in setting up the EMM QDB. It also offered the Task 9.2 team and ETHMIGSURVEYDATA with the additional resources needed to identify appropriate responses for each of the key challenges described in section 3, Identifying and overcoming the key challenges of creating a collection for the EQB dedicated to EMM surveys.

3. Identifying and overcoming the key challenges of creating a collection for the EQB dedicated to EMM surveys

As briefly discussed in the section above, Conceptualizing a collection for the EQB dedicated to EMM surveys, the key challenges for creating an EMM collection for the QDB were identified in parallel with determining how this service should be conceptualized. More precisely, the following five key challenges were identified:

- How to extract the text from a questionnaire
- How to document the questionnaires so that the corresponding metadata can be harvested
- How to handle the variety of languages used in questionnaires of EMM surveys
- How to create a public-facing interface for the EMM QDB
- How to integrate the EMM QDB into the EQB

All these challenges were originally confronted from the initial expectation that the previous work on the EMM Survey Registry would identify around 800 EMM surveys in total. However, the eventual compilation of the survey metadata uncovered nearly 1,700 surveys and, hence, this report takes into account this volume of surveys in the considerations around each challenge.

This section is thus structured, so that each subsection describes in detail an identified key challenge and explains which solution the ethnic and migration studies data community opted for to overcome the key challenge in question, including how they arrived at this decision.

3.1 Challenge 1: Extracting the text from a questionnaire

For both the EMM QDB and the EQB (where the EMM collection would exist), the text of the questionnaire needs to be extracted. This is because both services are aimed at providing users with an interface where they can discover and learn about existing questionnaires, by having access to their specific text-based questionnaire items (i.e. question wordings, code lists or response options/categories, statements/instructions, routing indications, etc.). As, initially, the expectation was to handle around 800 surveys (which could also be translated into multiple languages within any given location), there was a priori a strong preference to handle the inputting of the content of questionnaires through some form of automated process for efficiency purposes.

Thus, as a first step, the ethnic and migration studies data community examined roughly 450 surveys (i.e. the surveys detected after more than half of the countries had finished their scoping exercise of EMM surveys in their respective country) to determine how questionnaires were typically being documented and/or preserved. As the ethnic and migration studies data community found that most questionnaires were offered as a PDF or Word document, they leveraged the SSHOC network to identify experts who might be aware of existing technologies that allow for an autonomous or even semi-autonomous structured extraction of text found in PDF and/or Word documents, such that the

various elements included in any questionnaire could be automatically (or semi-automatically) be fed into a questionnaire and metadata documentation tool.

Most notably, the ethnic and migration studies data community connected with the leader of Work Package 4 on Innovations in Data Production, from the European Social Survey, Universitat Pompeu Fabra Barcelona (ESS-UPF). Using their expertise in translation practices in multilingual survey projects and corpora creation using machine translation, a training session was hosted at UPF in April 2019, where the arduous and complicated process of creating a workflow, using programs such as R,[17] for the automatic extraction of text from a single questionnaire, was demonstrated. As, at this training session, it became evident that there would be no efficient way to fully automate the process of structured text extraction, particularly in terms of the volume of questionnaires anticipated to be included in the EMM QDB, the following preliminary decision was made: to ensure the workflow for the EMM QDB included dedicated time to extract text manually.

The team later discovered that colleagues at the Institut national d'études démographiques (INED) in France, overseeing the documentation and dissemination of the Generations & Gender Programme (GGP) survey,[18] had successfully been using a semi-automatic process to extract structured text for questionnaires available in .docx format and formatted in a relatively uniform manner, using an R code that adapts for each individual questionnaire a generic publicly available code to get structured tables out of the .docx format document.[19] As such, the ethnic and migration studies data community met with the INED colleagues to examine whether their semi-automatic process could be replicated for the EMM QDB. With the support a data engineer for the CEE at Sciences Po, it was determined that the R code used by INED could be leveraged if the following steps were executed:

- **Step 1:** The .docx file is converted to a .xml file.
- **Step 2:** The .xml file is examined to identify the tags used to identify the various types of text (e.g. the question text, interviewer instructions, question name, etc.).
- **Step 3:** The R code from INED is adapted based on the tags identified in the .xml file.
- **Step 4:** The adapted R code is used to generate an Excel file that includes the question text, interviewer instructions, and question names.
- **Step 5:** The generated Excel file is imported into Colectica Designer, "a software for documenting data collections, survey specifications, and dataset descriptions" (Colectica, 2021b), producing a new version of the questionnaire displaying all the extracted text.

It is, however, important to note that this process, even if a questionnaire is available in .docx format and Colectica Designer is available for use, has a number of noteworthy limitations. First, certain types of questions, like grids,[20] cannot be properly extracted due to how they are formatted. Second, code lists (lists with the response categories) cannot be extracted, even if tags are used and found in the .xml file. Third, if the EMM QDB opts to use its own naming protocol for the question names, then all

[17] R is an open-source software for statistical computing and graphics. More information about R can be found at: https://www.r-project.org/ [11 November 2021]
[18] More information about GGP, including their role in SSHOC can be found on their website: https://www.ggp-i.org/ [11 November 2021]
[19] "Using R to get data out of Word docs": https://www.r-bloggers.com/2015/08/using-r-to-get-data-out-of-word-docs/ [10 December 2021]
[20] A question grid is one where the response options are exactly the same for a number of sub-items and, as a result, the question is displayed in some tabular form. For example, a question that asks respondents to state their level of confidence (on a 0-10 scale) in a number of different institutions and actors (e.g. politicians, parliament, the government, the police, the judiciary, etc.) is typically constructed as a grid.

question names need to manually be modified or updated. Fourth, even if tags are consistently used within the .xml file, the R code may still make errors during the extraction process (e.g. not extracting the interview instructions, identifying question text as interview instructions). Thus, as this process still necessitates a significant amount of human involvement to properly extract a questionnaire's text, the ethnic and migration studies data community determined that it should be used on a case-by-case basis, prioritizing questionnaires for which the semi-automatic extraction of text would result in significant time savings given the questionnaire length.

Finally, after finalizing the decision to use Colectica softwares for documenting the questionnaires and metadata of the EMM QDB (see section 3.2, Challenge 2: Documenting the questionnaires and metadata for the specifics), the ethnic and migration studies data community also learned from the INED colleagues that surveys (previously) documented in DDI Codebook and available on Nesstar—a metadata editor and publishing tool developed and offered by the Norwegian CESSDA SP, the Norwegian Centre for Research Data (NSD)—,[21] could have their corresponding .xml files imported into Colectica Designer if technical support could be provided by the Colectica central team. Furthermore, based on their own experiences of trialing this process, the INED colleagues advised that the text is generally extracted properly, requiring minimal monitoring and modifications by a human. As such, if this option could be realized, then EMM QDB would be able to use a three-pronged approach for extracting text: (1) fully manual extraction of text, (2) semi-automatic extraction of text leveraging the R code from INED, and (3) semi-automatic extraction of text by importing DDI Codebook .xml files downloaded from Nesstar.

3.2 Challenge 2: Documenting the questionnaires and metadata to facilitate harvesting

A key aspect of developing the EMM QDB is determining how to document the questionnaires and metadata in DDI Lifecycle. Hence, an immediate point of action taken was identifying and exploring the different technological solutions for documenting the questionnaires and metadata. By leveraging both the SSHOC network (e.g. CESSDA SPs, UPF-ESS) and the ETHMIGSURVEYDATA members affiliated with data archiving/preservation-oriented institutions (e.g. National Center of Social Research (EKKE), GESIS), the ethnic and migration studies data community was able to identify five different and potentially viable technological solutions.

The first was the **GESIS Questionnaire Editor**, a free web-based tool intended for documenting questionnaires so that they can be exported in DDI Lifecycle.[22] Initially conceived by GESIS so that questionnaires eligible for the EQB could be properly documented, it was an attractive tool in terms of cost, its adherence to the open software movement, and the underpinning impetus for its creation. As such, the ethnic and migration studies data community employed a German-speaking research assistant (RA) to trial the then available alpha version of the GESIS Questionnaire Editor. While this RA found the tool relatively easy to navigate and offering a number of essential features for documenting questionnaires of EMM surveys, they identified a number of key limitations: (1) the user interface, while indicating that it was available in English and German, was only partially available in English; (2) there was no clear strategy for identifying or grouping questionnaires belonging to the same survey or study;

[21] The official website for Nesstar can be accessed at: http://www.nesstar.com/ [11 November 2021]
[22] Access to the landing page of the GESIS Questionnaire Editor: https://multiweb.gesis.org/labs/apps/qeditor/ [11 November 2021]

(3) there was no option for classifying and tagging questions based on concepts; and (4) there were limited options in terms of how questions could be documented (e.g. not feasible to set up a response scale).[23] Given the development stage of the GESIS Questionnaire Editor, combined with the key limitations identified by the RA, the ethnic and migration studies data community determined that this tool was not an appropriate match for the EMM QDB work.

The next technological solution explored was **Archivist**, a free and open-source tool designed by the CLOSER team for documenting questionnaires from the CLOSER project. Archivist was also set up so that the main output format would be DDI Lifecycle using the questionnaire profile (Mills, 2021).[24] As a tool that had been successfully used to document a variety and significant volume of questionnaires from the CLOSER project, it was highly developed and user-friendly, it came with detailed documentation explaining how questionnaires could be documented using workflows established and used by the CLOSER team.[25] However, with the tool requiring and necessitating regular IT support (as it runs using Ruby on Rails[26]) and also not being able to support multiple languages for questions without ad-hoc programming development that was out of scope (and out of budget) for the ethnic and migration studies data community, it was determined that this tool was also unable to meet the questionnaire and metadata documentation requirements of the EMM QDB within the existing constraints and resources.

The remaining three technological solutions to consider were all Colectica softwares: **Colectica for Excel**, **Colectica Questionnaires**, and **Colectica Designer**.[27] First, Colectica for Excel was explored, as it was already being leveraged by the CDSP (also a SSHOC partner) and it was explained as a free software for documenting in DDI data structured in an Excel sheet. However, this tool was immediately ruled out as an option, as it produced DDI Lifecycle documentation using data matrices (i.e. not questionnaires, not even those preserved as Excel files). The ethnic and migration studies data community therefore evaluated Colectica Questionnaires and Colectica Designer, both paid softwares designed to handle the documentation of questionnaires in DDI Lifecycle. To do this, they liaised primarily with NSD, as they had been coordinating and developing the Questionnaire Variable Database (QVDB):[28] a database, based on and built using Colectica softwares, that allows users to explore relationships between and changes in questionnaires, questions (from questionnaires), and variables (a question's equivalent in the dataset) (SERISS, 2019). With support from NSD, the ethnic and migration studies data community learned how Colectica Questionnaires could easily be adopted to document questionnaires, as well as how Colectica Designer could be used to manage and document concepts for a questionnaire initially inputted via Colectica Questionnaires. Moreover, NSD explained that both Colectica Questionnaires and Colectica Designer have rich functionalities to adeptly capture different types of items found in questionnaires used for survey research. As a result of the knowledge gained

[23] The German-speaking RA produced an internal document detailing their experience in using the GESIS Questionnaire Editor. This document was subsequently used to identify the merits and limitations of this technological solution.
[24] Access to the GitHub page for Archivist and the CLOSER project website respectively: https://github.com/CLOSER-Cohorts/archivist and https://www.closer.ac.uk/ [12 November 2021]
[25] Access to the documentation describing the workflows for using Archivist: https://wiki.ucl.ac.uk/display/CLOS/Document+a+Questionnaire [12 November 2021]
[26] Ruby on Rails is an open source software that can be used for creating database-based web applications. It is offered under the MIT License. More information can be found on their website: https://rubyonrails.org/ [12 November 2021]
[27] Information about all products produced and made available by Colectica can be found on their website: https://www.colectica.com/software/ [12 November 2021]
[28] The QVDB was previously accessible at the following URL: https://colectica.nsd.no/. Due to ongoing technical development on this tool, the URL is currently inactive.

from NSD, it was determined that both Colectica Questionnaires and Colectica Designer could be used for documenting questionnaires and metadata for the EMM QDB.

Having selected Colectica Questionnaires and Colectica Designer as the ideal technological solutions, the team began the process of trialing these softwares using trial licenses. There were several significant obstacles that impeded the ability to trial Colectica Questionnaires and Colectica Designer in a stable and consistent manner. First, the need to identify dedicated research personnel to rigorously trial both softwares and fully vet that they could support the work of the EMM QDB[29]. Second, as Colectica Questionnaires and Colectica Designer operate strictly on Windows, the team needed to ensure that all parties involved in trialing these softwares would have access to a Windows-based computer[30]. Third, in order to use a Windows-based computer remotely to test the linkages between the Colectica Portal (for which the pre-production version[31] had been set up on the Sciences Po server) and Colectica Questionnaires and Colectica Designer respectively, VPN access needed to be authorized and configured. While authorization to use VPN access was swiftly approved, it required several trials, spanning roughly two months, to configure it properly. Fourth, as licenses for Colectica Questionnaires and Colectica Designer are on a per machine basis, they needed to determine how remote access to these softwares could be arranged for the eventual team of temporary RAs tasked with handling the documentation of questionnaires (and metadata) in specific languages. Given that Sciences Po was unable to set up this type of remote access (due to internal policies), they had to dedicate several months exploring alternative options; In September 2020, the team was able to establish a collaboration with Das Deutsche Zentrum für Integrations- und Migrationsforschung (DeZIM), a research institute participating in ETHMIGSURVEYDATA, where they could access a desktop containing licenses of both softwares by using Microsoft TeamViewer. Finally, trial licenses for Colectica Questionnaires and Colectica Designer could only be issued for short periods of time (1-3 months); as such, there were periodic disruptions in accessing the softwares as new trial licenses could not always be issued back-to-back, due to the working priorities of Colectica.

Colectica Questionnaires and Colectica Designer were rigorously tested to determine if and how these softwares could be jointly used to document questionnaires and metadata for the EMM QDB[32]. Despite also facing a steep learning curve in using Colectica Questionnaires and Colectica Designer, the ethnic and migration studies data community was able to finalize their choice of technical solutions as these two pieces of software. More importantly, they were able to undertake the preliminary work needed to develop the workflows for documenting questionnaires and metadata, as described in detail section 4.2, Developing and testing the workflows using the established thematic pilots.

[29] due to the health crisis (that is COVID-19), they were only able to recruit and onboard a post-doctoral researcher (funded by FAIRETHMIGQUANT) to undertake this work in September 2020

[30] due to the ongoing COVID-19 pandemic, during which the Sciences Po's information systems department has been facing increased demands in IT support, there were delays in securing Windows-based computers that could be used for any mandatory teleworking regulations (e.g. national lockdowns)

[31] The pre-production version of the CDSP's Colectica Portal is not publicly available. As such, it is intended for internal use to check that the questionnaires are properly documented and displayed, before making them available on the live version.

[32] This was possible once the aforementioned post-doctoral researcher began their post

3.3 Challenge 3: Handling the linguistic diversity of the questionnaires

For surveys including EMM respondents, it is common practice to offer the questionnaire in the mother tongues or native languages of the EMMs. In fact, as described in section 2.2, Discovering a need for a EMM Question Data Bank (QDB), the ethnic and migration studies data community discovered early on that roughly 50% of the EMM surveys with the questionnaire available for re-use offered it in at least one EMM-spoken language (see **Appendix: A**). Moreover, as many of the EMM surveys would not be fielded in an English-speaking country and/or with English-speaking respondents, it was expected that a sizable number of questionnaires would need to be translated into English (for the purposes maximizing their access and re-use via the EMM QDB).

Therefore, given this expected linguistic diversity of the questionnaires, the ethnic and migration studies data community began by exploring the available solutions for translating questionnaire text into English. They first leveraged the ETHMIGSURVEYDATA membership to understand how major and well-established international survey programs, like the ESS, handle the translation of their questionnaires; as they learned that professional translations, which typically use manual processes, are costly (both in terms of time and human resources needed), they shifted focus to determine if any automatic translation techniques could be utilized. Unfortunately, during the aforementioned training session at UPF in April 2019, they discovered that automatic translation would not be possible for the EMM QDB, as existing Natural Language Processing (NLP) translation tools are typically restricted to select languages (e.g. English, French, Spanish) and still require a human translator to rigorously quality check the translations they produce.

Having thus determined that the English translation work for the EMM QDB would need to be done manually, the next step for the ethnic and migration studies data community was to strategically decide the questionnaires requiring English translation. As such a decision would inevitably be informed by how the questionnaires and metadata would be documented (since any translated text would be part of the documentation process), the ethnic and migration studies data community was only able to develop a protocol for English translation once they selected Colectica Questionnaires and Colectica Designer as the actual tools for documenting the questionnaires and metadata. They thus completed a number of trials to see how different languages could be captured via Colectica Questionnaires and Colectica Designer and subsequently displayed on Colectica Portal. As a result, they determined that the English translation work would best be handled as follows:

A. If a non-English version of a questionnaire is a translation of the core/main questionnaire in English, then the non-English version would not need to be translated into English. Instead, the non-English version would offer an "English translation" by linking to the core/main questionnaire from which it was derived in the Colectica Portal.
B. If the non-English and English versions of a questionnaire were both used for fieldwork in the same location (e.g. English and French questionnaires were used for a survey conducted in London), then the non-English version would not need to be translated into English. Instead, the non-English version would offer an "English translation" by linking to the English version of the questionnaire used for fieldwork in the same location in the Colectica Portal.
C. If a non-English version of a questionnaire has no equivalent English version (i.e. does not fulfill scenarios A or B above), then it needs to be translated into English. The translated English text

would then be inputted and displayed alongside the non-English text as part of the questionnaire and metadata documentation process in Colectiona Questionnaires.

It should be noted that a sizable number of EMM surveys will have questionnaires that fall under scenario C above. As such, even if this English translation protocol is a realistic strategy for limiting the number of questionnaires to be translated into English, it is still contingent on whether the following human resources-related obstacles can be overcome: (1) being able to recruit and hire RAs who, on top of being able to properly document the questionnaires and metadata, have appropriate language skills to translate non-English text into English and (2) being able to establish stable remote access to Colectica Questionnaires and Designer, so that RAs with the desired aforementioned language skills can more easily be recruited[33] and, when ultimately hired, can work without disruption if any new COVID-19 work restrictions are put in place (e.g. work-from-home requirements). And, at time of writing this report, the ethnic and migration studies data community is actively exploring how to best address both obstacles.

Finally, because of the way in which the English translation protocol was ultimately defined, the ethnic and migration studies data community has been able to leverage this protocol as the overall logic for handling questionnaires with language variations. Specifically, Colectica Portal would show linkages for the following types of language variations: all questionnaires that are translations of the main/core questionnaire, all questionnaires used for fieldwork in a specific location, and all questionnaires used for a survey or study as a whole.

3.4 Challenge 4: Creating a public-facing interface for the EMM QDB

The public-facing interface of the EMM QDB is the online space where users would be able to access and exploit the questionnaires and their questionnaire items. Once the idea of the EMM QDB was born, an immediate decision was made to not build this service from scratch, but to use and leverage an existing web-based application for handling questionnaires; this decision was heavily informed by the experience of contracting an IT company to fully design the EMM Survey Registry (i.e. D9.4), as the technical and resource demands were high, despite the service being relatively less complex than that of the envisioned EMM QDB. Moreover, once it was determined that the questionnaires and metadata would need to be documented using both Colectica Questionnaires and Colectica Designer, it became immediately evident that the public-facing interface would ideally need to be based on a Colectica Portal.[34] This is because Colectica Portal is the web-based application created by Colectica to display data (including questionnaires) and the corresponding metadata, meaning that it was set up to seamlessly ingest and publish questionnaires and metadata documented using Colectica Questionnaires and/or Colectica Designer (Colectica, 2021a).

Thus, having identified Colectica Portal as the ideal web-based application to use, the ethnic and migration studies data community determined and explored two main options for accessing and using

[33] For certain EMM-spoken languages (e.g. Arabic, Hungarian, and Mandarin), it can be challenging to find individuals who speak these languages plus English and can also work in Paris/France (i.e. the location where the Task 9.2 team is based). Therefore, by allowing remote access to Colectica Questionnaires and Designer, it could improve access to individuals with diverse language skills, as they could work while being based outside of Paris/France.

[34] Information about Colectica Portal made available on the Colectica website: https://www.colectica.com/software/portal/ [11 November 2021]

Colectica Portal. The first would be purchasing and setting up their own Colectica Portal strictly for the EMM QDB. The second would be using a Colectica Portal belonging to a CESSDA SP. Ultimately, the second option was deemed the best solution for the following main reasons:

- Purchasing Colectica Portal is expensive, subsequently leaving limited funds for hiring research staff needed to actually develop the EMM QDB (e.g. RAs to document the questionnaires and metadata using Colectica Questionnaires).
- SSHOC, FAIRETHMIGQUANT, and ETHMIGSURVEYDATA are all projects, meaning that an EMM QDB-dedicated Colectica Portal could only be assured sustainability for the lifespans of these projects bringing too much uncertainty into long-term sustainability.
- The EQB's first priority is to curate questionnaires by harvesting metadata for surveys held by CESSDA SPs, hence harvesting metadata from non-CESSDA SP portals would present additional difficulties.

As such, the ethnic and migration studies data community liaised with the CDSP (a SSHOC and FAIRETHMIGQUANT Sciences Po-based partner) to determine if their Colectica Portal could be used for the EMM QDB and if the plans included long-term sustainability of this option. After having successfully set up the pre-production version of their Colectica Portal, the CDSP formally confirmed that it would be feasible and that their own strategic and financial planning foresaw sustaining their own Colectica Portal at least for the next 10 years. As a result, the the ethnic and migration studies data community was able to secure a solution, where the public-facing interface of the EMM QDB would be set up in a durable, sustainable, and EQB-compatible manner. These presented many advantages (notably, medium-term sustainability and seamless compatibility with EQB workflows), but it also presented the disadvantage that the ethnic and migration studies data community would not be in full control of the timeline for the development and publication of the CDSP Colectica Portal that would host the EMM QDB.

3.5 Challenge 5: Integrating the EMM QDB into the EQB

As discussed previously, the EQB's priority has been to curate and ingest questionnaires for surveys held by CESSDA SPs. Consequently, in order to facilitate the integration of the EMM QDB into the EQB (and subsequently set up the EMM collection), the ethnic and migration studies data community would ideally need to identify a way to set up the EMM QDB with a CESSDA SP, so that the metadata for the questionnaires of EMM surveys can be harvested by the EQB.

Some of the options that were initially considered (e.g. setting up an independent Colectica Portal exclusively dedicated to the EMM QDB, or one hosted by other collaborators such as DeZIM) were suboptimal because they would depart from the preferred and pre-established workflows for EQB that relied on harvesting (as a matter of priority) the CESSDA SPs. Fortunately, the CDSP was able to confirm that it could support and include the EMM QDB as part of their Colectica Portal as mentioned above. Moreover, as the CDSP is a participating partner in SSHOC and FAIRETHMIGQUANT, it would be able to actively liaise and rigorously test with CESSDA MO the process of harvesting the EMM QDB's questionnaires and metadata.

Due to technical delays resulting from the ongoing COVID-19 pandemic, the live version of the CDSP's Colectica Portal was only released in early December 2021.[35] Consequently, the CDSP has not yet been able to provide a valid Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) endpoint, which serves as the single access point from which the EMM QDB's metadata (documented in DDI Lifecycle) can be harvested by the EQB. Nevertheless, the CDSP is still strongly positioned to facilitate and ensure the integration of the EMM QDB into the EQB and that completing the next steps is only a matter of time.

[35] Access to the live version of the CDSP's Colectica Portal: https://explore.cdsp.sciences-po.fr/ [3 December 2021]

4. Determining the workflows for the EMM QDB

In order to develop the actual EMM QDB, the ethnic and migration studies data community needed to determine the specific workflows to use. As such, this section begins by describing why the decision was made to create and use thematic pilots to develop and test the workflows for setting up the EMM QDB. The section then details the workflow development processes, which were undertaken via the thematic pilots and heavily informed by the solutions identified and adopted for the key challenges described in 3: Identifying and overcoming the key challenges of creating a collection for the EQB dedicated to EMM surveys).

4.1 Establishing thematic pilots

Opting for a pilot-based approach for developing the EMM QDB, the main rationale was that, by restricting the number, focus, and types of questionnaires to be handled, it would be possible to rigorously test and vet the workflows to be used, as well as more efficiently handle the stages relating to conceptual hierarchical mapping of questionnaire items (as described in the paragraphs below).

Consequently, the ethnic and migration studies data community determined that the EMM QDB pilots (n=3) would be centered on specific themes or topics, based on the topic classifications of the EMM Survey Registry (i.e. variable 1.13 of the EMM Survey Registry's metadata schema),[36] as well as the main areas of expertise of the ETHMGISURVEYDATA membership. More precisely, it was determined that the first pilot would cover the topic of civic life and political integration (variables 1.13.20, 1.13.24, and 1.13.27); then once a clear workflow was established for this initial topic, pilots for the following two additional topics would be carried out: stereotypes, prejudices, discrimination, intergroup relations/beliefs (variables 1.13.5, 1.13.14, and 1.13.21); and health access (variable 1.13.10). Additionally, a second important decision was made to have these three thematic pilots cover nine of the 35 countries formally participating in ETHMIGSURVEYDATA: Croatia, Germany, Hungary, Italy, the Netherlands, Norway, Spain, Switzerland, and the UK; these 9 countries were strategically selected for their geographic and linguistic diversity, as well as their generally differing survey research landscapes.

To facilitate the roll out of these three thematic pilots, ETHMIGSURVEYDATA, with coordination support from the Task 9.2 team, focused on finishing the metadata compilation process for EMM surveys of the nine pilot-participating countries. Moreover, ETHMIGSURVEYDATA established internal task forces for each topic, where each task force would be responsible for proposing a conceptual hierarchy mapping for their corresponding topic in connection to dimensions of integration and/or inclusion; the task force's proposed conceptual hierarchy would then be used to map and show relationships for questions (from the questionnaires of various EMM surveys) belonging to the same topic(s).

While the pilot-participating countries were all able to compile metadata for the EMM surveys detected in their respective country (albeit with slight delays for select countries, due to resource constraints), the task forces faced repeated delays in developing their respective conceptual hierarchy, due to the high intellectual and time demands of this type of work in the context of ETHMIGSURVEYDATA (which

[36] The Excel-based version of the EMM Survey Registry's metadata schema has been published to and made available via Zenodo: https://zenodo.org/record/4676947 [13 November 2021]

was undertaken on a voluntary basis and with no person-month funding). In fact, the task force for health access determined that they could no longer deliver a conceptual hierarchy, resulting in a replacement thematic pilot to be identified: belonging and social identity, cultural and social norms (variables 1.13.8, 1.13.12, and 1.13.20). The identification of this alternative third pilot focus was done through active exchanges with a range of stakeholders (notably EU-level policy-makers and policy advisors) during the third Policy Dialogue conference organized by ETHMIGSURVEYDATA on 9-10 March 2020 at Brussels.[37] A number of stakeholders expressed that, given the considerable availability of data resources on other dimensions of integration (such as employment, education, etc.), the major gaps were to be found on access to data and indicators relating to 'soft' integration indicators that are not readily available from official statistics and trans-European large-scale surveys.

Fortunately, at the time when the new thematic pilot was established, the FAIRETHMIGQUANT project had already gone into effect. With one of the main commitments of the FAIRETHMIGQUANT project being to develop and test the workflows for the EMM QDB via the thematic pilots and specifically for EMM surveys conducted in France, the ethnic and migration studies data community was able to leverage the FAIRETHMIGQUANT resources to provide enhanced coordination support to the task forces. More importantly, with FAIRETHMIGQUANT being a project funded by the French ANR, the thematic pilots were refined for a final time to include France as a participating country.

Thus, with the parameters of the thematic pilots finalized, the work of developing and testing the workflows for the EMM QDB began, under the leadership of the FAIRETHMIGQUANT team. The section below describes in more detail how the workflows to use came to fruition via the thematic pilots.

4.2 Developing and testing the workflows using the established thematic pilots

The process of setting up the EMM QDB can be broken up into two major and sequential components: first, the retrieval of questionnaires for the EMM surveys and second, the documentation of the retrieved questionnaires and their corresponding metadata. This section is, therefore, organized into subsections, where each subsection describes the evolution of the workflow for the corresponding component, resulting from the trials conducted via the aforementioned established thematic pilots (civic life and political integration; stereotypes, prejudices, discrimination, intergroup relations/beliefs; belonging and social identity, cultural and social norms). Moreover, each subsection highlights the major points of decision that were critical for developing the final workflow.

4.2.1 Retrieving questionnaires

Prior to defining the thematic pilots, the ethnic and migration studies data community determined that the EMM QDB should have generous inclusion criteria to maximize the number of questionnaires covered by the EMM QDB. In others words, they agreed to retrieve a questionnaire for the EMM QDB, regardless of its quality, as long as it was publicly available (i.e. is available online and can be

[37] Access to the meeting agenda and session recordings for the ETHMIGSURVEYDATA Policy Dialogue conference in Brussels on 9-10 March 2020 respectively:
https://ethmigsurveydatahub.eu/wg-plenary-meeting-and-policy-dialogue-conference/ and
https://www.youtube.com/playlist?list=PLo7wmLY6eAWLhrPToIVJeEc_3NKhD4XMm

downloaded, even if an account must be created to do so) or accessible through a member of ETHMIGSURVEYDATA.

With this inclusion criteria for the EMM QDB in mind, the ethnic and migration studies data community proceeded by exploring GDPR-compliant file storing services that could be used to store all questionnaires eligible for the EMM QDB. They considered a number of EU-developed services, like B2DROP by EUDAT,[38] but ultimately settled on MyCore by the Centre national de la recherche scientifique (CNRS) in France.[39] The principal reasons for selecting MyCore were that this service had been developed by a reputable and trusted entity (as the CNRS is who the French state has entrusted to coordinate scientific research in France) and had demonstrated good usability, when testing using select questionnaires qualifying for the first thematic pilot (civic life and political integration).

Also, in parallel with determining where to store the questionnaires for the EMM QDB, the ethnic and migration studies data community developed an initial protocol (see **Appendix: B**) for identifying, obtaining, and storing eligible questionnaires for the first thematic pilot, leveraging the already live beta version of the EMM Survey Registry. Their initial protocol, which was tested by a team of RAs funded by SSHOC and/or FAIRETHMIGQUANT, consisted of the following steps:

- **Step 1:** Using the live version of the EMM Survey Registry, filter for surveys that: (1) have been conducted in at least one of the ten pilot-participating countries (variable 1.0), (2) cover at least one of the civic life and political integration variables (variables 1.13.20, 1.13.24, and 1.13.27), and (3) have been identified as having its questionnaire(s) either publicly available or available via a COST Action member (variable 8.16).
- **Step 2:** For each survey identified, as part of the filtering process described in step 1 above, obtain all of its available questionnaires using the provided URLs or by contacting the relevant COST Action member (variable 8.16).
- **Step 3:** For each obtained questionnaire, store it on the MyCore space set up for the EMM QDB (which is only accessible to authorized users), using the established file organization and naming protocols.
- **Step 4:** For each survey identified, fill out an internal tracker to document essential information about all of the questionnaires obtained (e.g. which survey the obtained questionnaires belong to, when and who obtained and stored (onto MyCore) the questionnaires, which language(s) the obtained questionnaires are in).

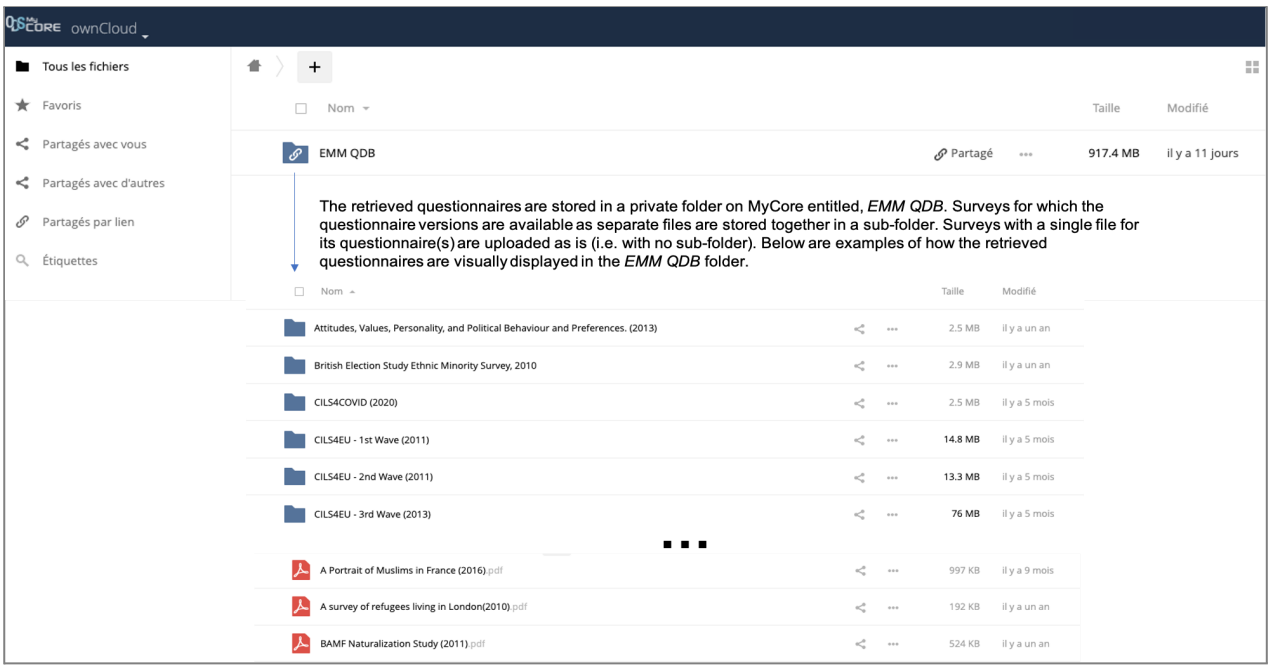
Overall, the RAs were able to implement and execute the initial protocol without major issues. In fact, the only step that required modification was step 3, in relation to the naming protocol; specifically, a decision was made to adapt the naming protocol so that questionnaires belonging to a larger survey or study can more easily be identified and located within MyCore.

Thus, a workflow for retrieving questionnaires for the EMM QDB was ultimately established, due to the successful trials conducted via the first thematic pilot. The established workflow was then repeatedly re-tested and further validated, whenever new surveys from the pilot-participating countries were added to the EMM Survey Registry; in fact, at the time of writing this report, the established workflow has been used to screen almost all the surveys to be considered for the first thematic pilot,[40] leading to

[38] Access to the user documentation for B2DROP: https://eudat.eu/services/userdoc/b2drop [12 November 2021]
[39] Landing page of MyCore: https://mycore.core-cloud.net/index.php/login [12 November 2021]
[40] As of December 2021, there are 17 surveys that need to be checked to see if they qualify for the first thematic pilot.

413 questionnaires (which includes all variations of a given questionnaire) to be successfully uploaded to MyCore (see **image 1**). However, it is important to note that the established workflow has not yet been applied to any of the other thematic pilots; as explained in more detail in the section below, the two other thematic pilots have yet to be deployed, given the immense time and cognitive demands that have been required in determining the workflow for documenting the questionnaires and metadata.

Image 1: Example of how questionnaires are stored in MyCore



4.2.2 Documenting the questionnaires and metadata

With the questionnaires and metadata to be documented using Colectica Questionnaires and Colectica Designer, the team connected with NSD to deepen their understanding of how they have been actually leveraging these softwares to document questionnaires and metadata for the QVDB. From these exchanges, it became increasingly evident that Colectica Questionnaires, given its user-friendly and intuitive interface, should be used for documenting the questionnaire and questionnaire items, whereas Colectica Designer should be used for functionalities that are not available in Colectica Questionnaires, e.g., managing and documenting concepts (identified as part of the conceptual hierarchy work of the ETHMIGSURVEYDATA task forces) and linking across different questionnaire versions and question items from different questionnaires.

As a next step, FAIRETHMIGQUANT did a first attempt of documenting questionnaires from the first thematic pilot (i.e. select questionnaires from the LOCALMULTIDEM survey)[41] using Colectica

[41] LOCALMULTIDEM is a survey studying the degree of political integration of migrants or foreign-oirign persons in several European cities. It is also a survey that was led by Laura Morales. More information about LOCALMULTIDEM can be found on their Harvard Dataverse page: https://dataverse.harvard.edu/dataverse/localmultidem#:~:text=Sort-,LOCALMULTIDEM%20and%20MDE%20Individual%20Survey%20(WP4)%20Dataset%2C%202004-2008,-Apr%202%2C%202017 [12 November 2021]

Questionnaires and leveraging the techniques shared by NSD. After documenting randomly selected fragments of the LOCALMULTIDEM questionnaires, they imported these partially documented questionnaires into Colectica Designer to first trial the process of adding the conceptual information to the question items (a Colectica controlled vocabulary to refer to the question (regardless of type), including all the information attached to it, such as the code list or interviewer instructions/statements). While the question items documented generally appeared properly when viewed in Colectica Designer and could be assigned specific concepts, there was one notable exception: question grids. Specifically, it became evident that question grids would need to be documented as a sequence of questions, as opposed to using the pre-set question grid option provided by Colectica Questionnaires; this is because when the pre-set question grid option is used, individual rows within the grid cannot be assigned specific concepts. This decision then also meant that the NSD protocol for establishing each question item as a sequence (which is advantageous for visualization purposes) could not be used, as there would be no way to quickly distinguish question grids.

The partially documented questionnaires, imported into Colectica Designer, were also able to be leveraged to test how questionnaire versions, as well as question items from different questionnaires could be linked with one another. To understand the different ways in which such linkages could be represented, they consulted Colectica and the CLOSER project, but also explored Colectica Portals of other research groups/projects (e.g. QVDB, CLOSER Discovery,[42] CDSP's pre-production version). However, despite testing the different linkage options, it was difficult to determine the ideal options for the EMM QDB without being able to see the outputs of these tests on the CDSP's Colectica Portal. As such, they accelerated access to the CDSP's pre-production Colectica Portal (the only version released) for both herself and the other members of the Task 9.2 team.

Once stable access to the CDSP's pre-production Colectica Portal was established, it was possible to publish the partially documented LOCALMULTIDEM questionnaires (as well as other "tests'' using questionnaires from the first thematic pilot). This then allowed them to trial various naming conventions, display arrangements, and methods for linking questionnaire versions and question items from different questionnaires. After conducting a series of trials over the course of a year (which were all carefully reviewed and analyzed by the ethnic and migration studies data community), the following key decisions were made:

- **Documenting question items:**
    - Colectica Questionnaires would be used to document all question items found in the questionnaire. In rare instances where a question item cannot be documented fully within Colectica Questionnaires (e.g. a special code list needs to be inputted), Colectica Designer would be utilized.
    - Only question grids would continue to be documented as a sequence of questions. All other types of questions would be adapted to use the pre-set options of Colectica Questionnaires for efficiency.
    - If a questionnaire is to be documented by inputting and using text resulting from one of the two semi-automatic text extraction processes (see section 2.1, Challenge 1: Extracting the text from a questionnaire, for specifics), Colectica Questionnaires and Colectica Designer would jointly and strategically be used to minimize the time spent to edit, modify, and enhance the extracted text.

[42] Access to the landing page of CLOSER Discovery: https://discovery.closer.ac.uk/ [12 November 2021]

- **Conceptual hierarchies and their concepts:**
  - The conceptual hierarchy would be defined by concepts and sub-concepts; the concepts and sub-concepts belonging to the same conceptual hierarchy would be grouped into concept sets using Colectica Designer. The hierarchy of concepts and sub-concepts would also be established by denoting a sub-concept as a subclass of the superior concept.
  - Question items would be tagged, using Colectica Designer, with all relevant concepts (see **images 2-4**).[43]
- **Linking question items across questionnaires (see images 2-4):**
  - Both represented and conceptual variables would be used to link across different question items that are different versions of the same question. Linkages would all be established using Colectica Designer.
  - Each question item would be described by a represented variable that can be uniquely identified due to the naming conventions used for its name and label (see bullet point below on naming conventions).
  - Conceptual variables would be added for each version of the question within the overall study. They would only be labeled (no names) using information on the study and concept (see bullet point below on naming conventions).
  - Represented variables would be displayed as sub-items of the broader conceptual variable and a correspondence tree within the Portal allows for the visualization of and linking to related question items.
- **Linking questionnaire versions:** Different questionnaire versions (i.e. those belonging to the same survey or study, including those that are language variations of one another) would be linked using Colectica Designer. These linkages would then be visible in the visualization trees within Colectica Portal, where all the different questionnaire versions would be identified as different instruments of the given survey or study (see **image 5**).
- **Naming conventions (see images 2-4, 6-9):**
  - Question names would be written to contain information on the survey/study, location, language, and question ID/number. The question name would specifically be written as: [Survey/study acronym]-[Study territorial location]-[Language in ISO 693-3 code]-[Question ID/number][44]
  - Question labels would be used to display conceptual information for the relevant topic. The question label (which would also mention the survey/study, location, and language) would be written as: Concept name (Study/survey acronym-Study territorial location-Language in ISO 693-3 code)[45]

[43] In relation to questionnaires that are documented as part of the thematic pilots, question items that do not correspond to any concept from the conceptual hierarchies of the thematic pilots, would not be tagged with any concepts.

[44] Here are two examples of how to write the question name. For the LOCALMULTIEM Lyon questionnaire, the first question (i.e. Q1) is a grid, with grid items identified using A, B, C, etc. As such, for the first grid item in Q1, the question name would be written as: LMD-Lyon-FRA-Q1_A. Another example is from the LOCALMULTIDEM master questionnaire. For the eighth question (i.e. Q8), the question name is written as LMD-Master-ENG-Q8.

[45] In relation to questionnaires that are documented as part of the thematic pilots, question items that do not correspond to any of the thematic pilot variables would not have any question labels yet. As a question label is not a mandatory field, this presents no functional problems. Within the Colectica Portal, the question name (which follows the same convention regardless of whether a question corresponds to a thematic variable) will be displayed and is sufficient to uniquely identify the question item and understand the questionnaire source and language. As more thematic variables are added, corresponding question labels will also be added at the same time as the conceptual mapping is done.

- Represented variables would be named to include information on the survey/study, location, and question ID/number only and would be written as: [Survey/study acronym]-[Study territorial location]-[Question ID/number]
- Labels for represented variables would contain information on the survey/study, location, and conceptual information. If applicable, it would also include other identifying information. The label would be written as: Question concept (Survey/study acronym - Study territorial location - (Other identifying information[46]))
- Conceptual variables would only be labeled (i.e. no name). The label for conceptual variables would contain the name of the specific concept and information on the study. The label would also be written as: Concept name (Study acronym)

The results of these various trials have subsequently been translated into a clear workflow (though a working version), which is documented in an internal Google Doc. The current version of the workflow (see **Appendix: C**) is now being used and an additional member of the Task 9.2 team to document, from scratch, a number of high priority or high interest questionnaires from the first thematic pilot (see **images 6-9** for examples of how documented questionnaires would appear on Colectica Portal). Thus far, the workflow has worked well and has also allowed the additional member of the Task 9.2 team to more quickly learn the intricacies and functionalities of Colectica Questionnaires and Colectica Designer.

To conclude, the process of establishing a workflow was an arduous and lengthy process (an estimated 3 person-months at FTE had to be devoted), given the steep learning curves of Colectica Questionnaires and Colectica Designer. The effort to develop a workflow was also highly impacted by the fact that there was no existing workflow to replicate or recycle[47] and that access to the software was routinely disrupted due to resource constraints, delays in having continuous access to the softwares (due to licensing issues), and technical compatibility between varying versions of the softwares (e.g. between Colectica Portal, Colectica Designer and Colectica Questionnaires). Nevertheless, given the numerous rigorous trials that were conducted to realize a workflow, as well as the recent success to test the current workflow with additional team members and questionnaires, the ethnic and migration studies data community is confident that the current workflow can be adopted to document new questionnaires, with limited modifications to the process needed.

[46] Other identifying information may be necessary when a question is asked in different versions for specific subpopulations but the questionnaire has designated these each as separate questions. An example is items H, I and J on the Q1 grid for LOCALMULTIDEM Lyon questionnaire. These ask about respondents' attachment to Algeria, Tunisia, and Morocco respectively. These questions all correspond to the concept "Self identification with country of origin or ancestry." Thus, to uniquely identify this question label within the study questionnaire, the label is designated as "Self-identification with country of origin or ancestry (LMD-Lyon-Algeria)" where Algeria provides the additional identifying information.

[47] For the EMM QDB, the ethnic and migration studies data community is only documenting the questionnaires and corresponding metadata. However, other users of Colectica Questionnaires and/or Colectica Designer are documenting questionnaires and the corresponding metadata, alongside the data matrices. As such, the documentation process when data matrices are used is different and cannot be exactly replicated or re-used for the EMM QDB.

Image 2: Display of question items linkages across different questionnaires on Colectica Portal

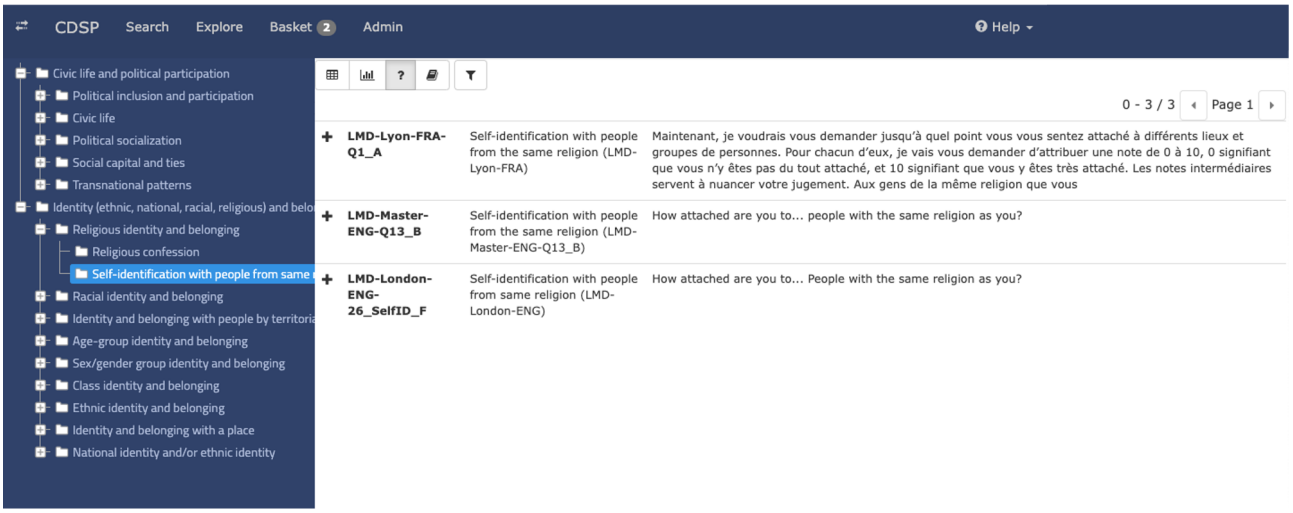


Image 3: Display of a conceptual variable on Colectica Portal

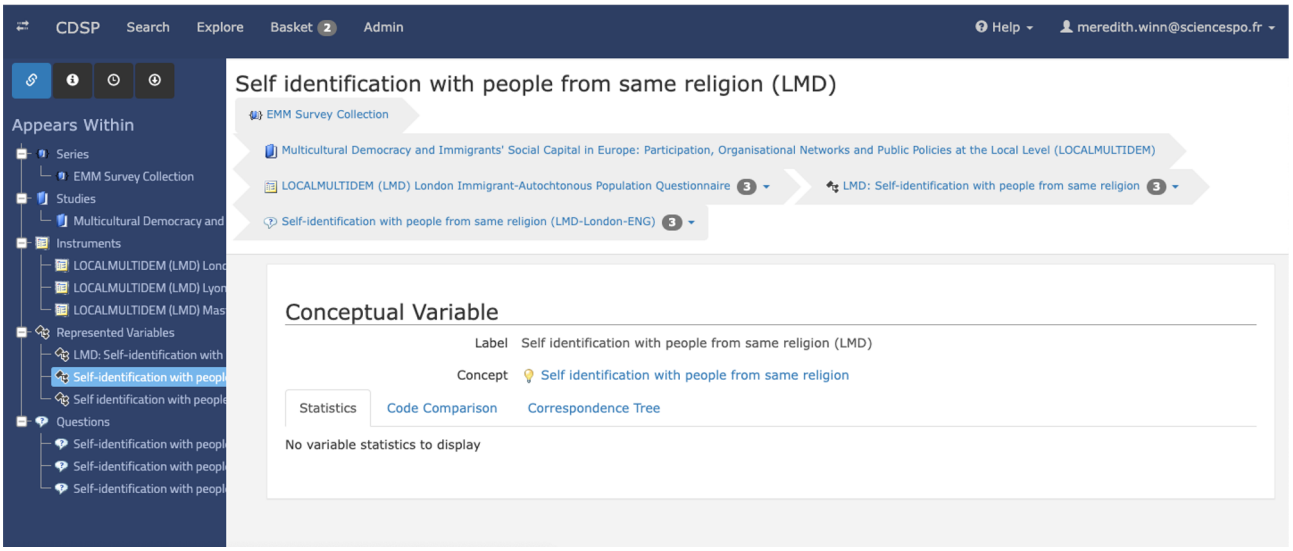


Image 4: Annotated display of a conceptual variable on Colectica Portal

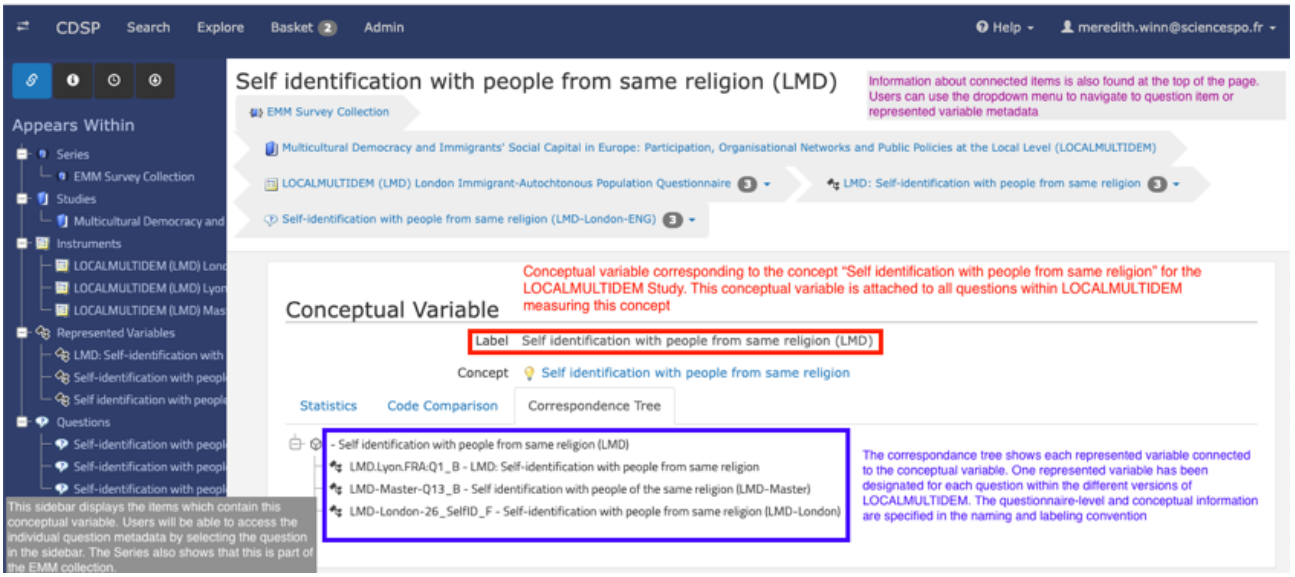


Image 5: Display of different questionnaires belonging to the same larger survey or study on Colectica Portal

Image 6: Example of a properly documented questionnaire displayed on Colectica Portal (part 1)

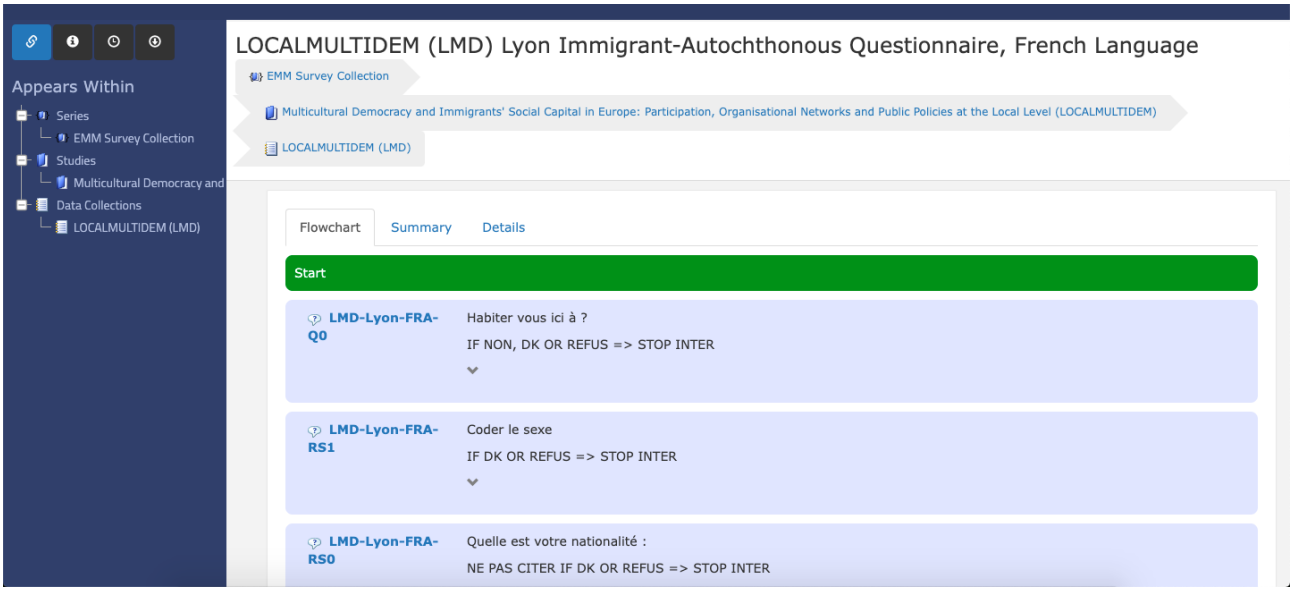


Image 7: Example of a properly documented questionnaire displayed on Colectica Portal (part 2)

Image 8: Example of a properly documented question item displayed on Colectica Portal (part 1)

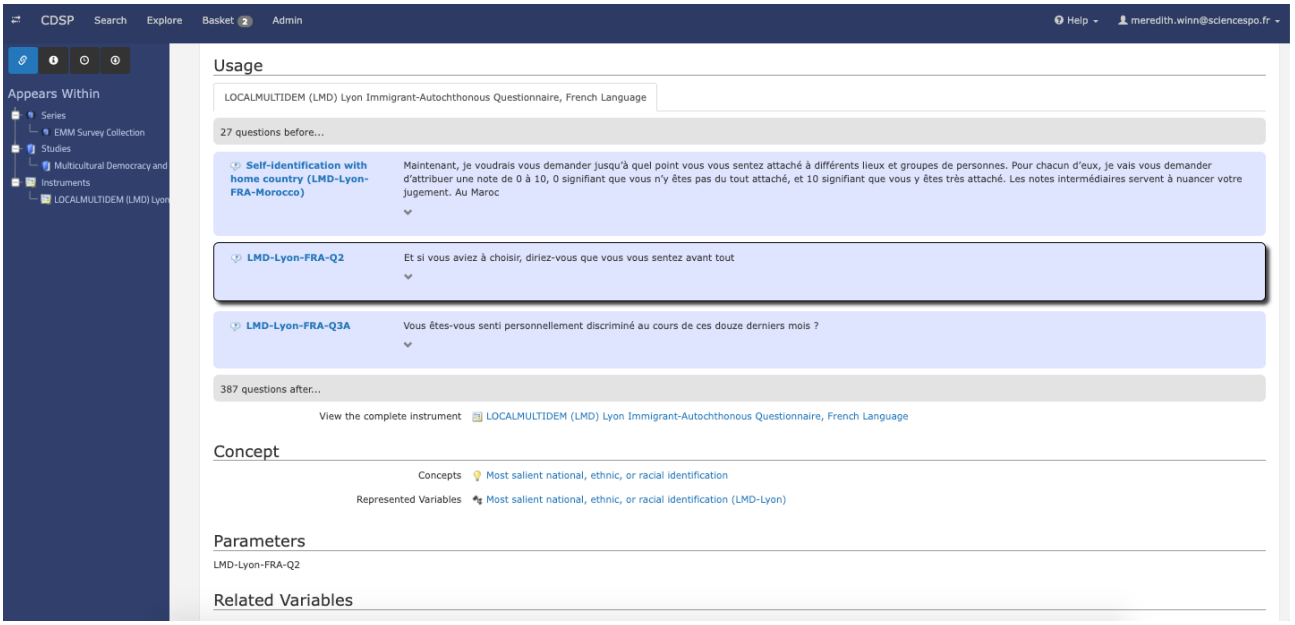
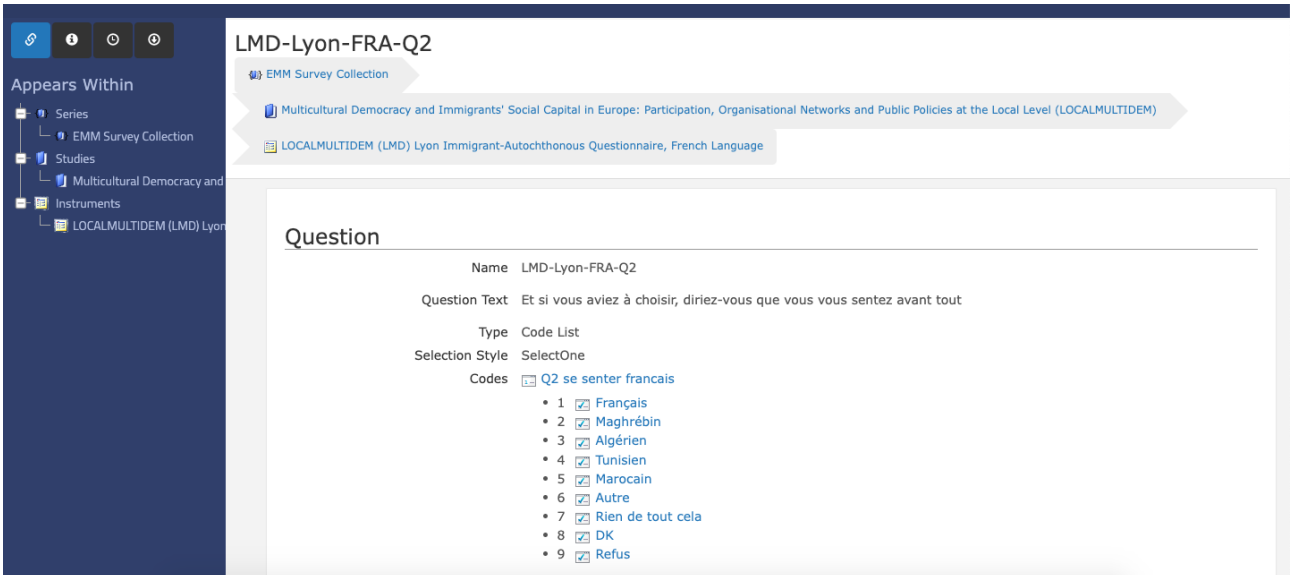


Image 9: Example of a properly documented question item displayed on Colectica Portal (part 2)



5. Linking the EMM QDB to the EQB

The full scale of the EMM QDB was originally conceived to cover all EMM surveys, identified as part of the EMM Survey Registry work, for which the questionnaire would be publicly available or available via a member of ETHMIGSURVEYDATA. Nevertheless, within the scope of the resources provided to the ethnic and migration studies data community within SSHOC, as well as the additional resources obtained through FAIRETHMIGQUANT, it was always clear that only piloting work would be feasible, before the service and tool could be up-scaled to include all the target questionnaires. As a result, in order to successfully link the EQB to what eventually could become (if continued funding support is available from EU and national open science and open data programs) the full scale EMM QDB—which would be created using the workflows described in section 4.2, Developing and testing the workflows using the established thematic pilots—the EMM QDB would need to be set up on the CDSP's Colectica Portal, so that its metadata can be harvested by the EQB. This section, therefore, explores the steps taken to ensure that the CDSP's Colectica Portal can host the EMM QDB. It also discusses how the EQB can properly harvest the metadata for the EMM QDB (from the CDSP's Colectica Portal), which is needed for creating the EMM collection of the EQB.

5.1 Setting up the EMM QDB on CDSP's Colectica Portal

The questionnaires included in the EMM QDB are to be documented using Colectica Questionnaires and Colectica Designer. As the CDSP's Colectica Portal is also a Colectica product, there are mechanisms already in place that allow questionnaires and metadata documented in Colectica Questionnaires and Colectica Designer to seamlessly be published to Colectica Portal (Colectica, 2021a). Nevertheless, in order to ensure such a linkage can be set up successfully, the following steps need to be executed:

1. The CDSP, as the managing party of the CDSP's Colectica Portal, needs to issue personal accounts for every user needing to publish to their Colectica Portal. These issued personal accounts can only be used if on the Sciences Po network (i.e. through VPN access and a Sciences Po email user account).
2. The ethnic and migration studies data community needs to ensure that their copies of Colectica Questionnaires and Colectica Designer can be synchronized with the CDSP's Colectica Portal. More specifically, they need to configure their copies to communicate with the Colectica Repository that is managing the CDSP's Colectica Portal. This configuration can be achieved by designating the Colectica Repository to use (via the unique identifier of the relevant Colectica Repository) and using the personal accounts issued by the CDSP (as described in point 1 above).

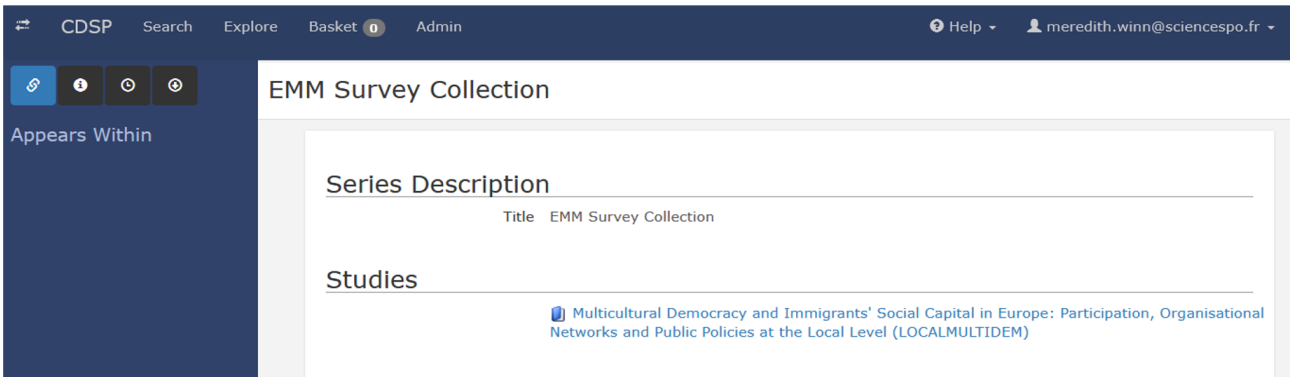
As the CDSP was able to set up the pre-production version of their Colectica Portal (i.e. the version that is internal and used to check that the questionnaires are documented and displayed properly), the ethnic and migration studies data community was able to use this version of the CDSP's Colectica Portal, later on in the year, to trial the two-step process described above. What they found is that the two steps can easily be implemented and, more importantly, once both steps are executed, questionnaires can instantaneously be published and become visible.

Moreover, the ethnic and migration studies data community was also able to leverage the pre-production version of the CDSP's Colectical Portal to establish how to best designate the

questionnaires of the EMM QDB as a separate collection from questionnaires being contributed via the CDSP. As Colectica Portals cannot use DDI agencies (i.e. unique identifiers created by DDI to designate the specific entity responsible for the production of the given metadata) or other unique identifiers to automatically build a dedicated collection, the decision was made to create a series in Colectica Designer entitled, EMM Survey Collection, where all documented questionnaires for the EMM QDB could be linked to or designated as being part of this series via Colectica Designer. The output of this strategy, as validated via the pre-production version of the CDSP's Colectical Portal, is that the EMM Survey Collection would be displayed as a unique collection, where all questionnaires belonging to the EMM Survey Collection would subsequently be made accessible (see **image 10**).[48]

Despite the aforementioned strides in understanding how the EMM QDB can be linked to the CDSP's Colectica Portal, at the time of writing this report, the live version of the CDSP's Colectica Portal was only configured and released in early December 2021 due to delays unexpectedly imposed by the ongoing COVID-19 pandemic. As such, one key question remains to be answered: Will the workflow for publishing tested and confirmed via the pre-production version also work for the live version? Thus, once the live version becomes available, the ethnic and migration studies data community will need to undergo a new round of testing to ensure that the EMM can successfully and sustainably be linked to the live version of the CDSP's Colectica Portal. Equally, the hierarchy currently proposed for display—the EMM Survey Collection as a series, studies as the individual survey projects within that series, and instruments as the individual questionnaires within each study—will also need to be further vetted, because there are still uncertainties as to whether it will allow for clear and user-friendly handling of cross-sectional repeated and longitudinal studies (e.g. Understanding Society in the UK, the European Social Survey, or the GGP surveys).

Image 10: The EMM QDB on the CDSP's pre-production Colectica Portal



5.2 Harvesting the EMM QDB's questionnaires for the EQB

With the EMM QDB to be hosted on the CDSP's Colectica Portal, the question of integrating the EMM QDB into the EQB rests on whether the EQB can successfully harvest metadata from the CDSP's Colectica Portal. More specifically, this means that the CDSP's Colectica Portal needs to be able to

[48] This solution is currently being trialed as it is yet unclear how easily this will then be identified as a separate collection within the published CDSP Colectica Portal and, importantly, how it will be displayed as a unique collection within EQB.

provide a valid OAI-PMH endpoint (i.e. a single access point for harvesting the metadata in DDI Lifecycle), as well as ensure that the questionnaires belonging to the EMM QDB are properly being documented in DDI Lifecycle (i.e. using DDI-Lifecycle XPathes that are compatible with the metadata schema used by the EQB) (Akdeniz, 2021) and correctly being displayed as part of a single and identifiable collection stemming from the ethnic and migration studies data community.

Unfortunately, at the time of completing this report, the ethnic and migration studies data community has been unable to test yet whether the EQB can harvest the EMM QDB metadata from the CDSP's Colectica Portal. This is due to a number of technical delays resulting from the ongoing COVID-19 pandemic.

First, due to resource constraints, which were further exacerbated by COVID-19, CESSDA SP GESIS who originally led the development of EQB (Krämer & Akdeniz, 2020), was unable to continue the development of EQB beyond the launched alpha version. Given this change, CESSDA ERIC began re-evaluating possibilities for offering a sustainable EQB and, following the decision and procurement process, it was not until November 2021 that they were able to re-configure the EQB using Colectica and offer an initial round of testing for harvesting the metadata.

Second, while the CDSP was able to swiftly set up their pre-production version of Colectica Portal, they have faced ongoing challenges and delays in releasing the live version, which is where a valid OAI-PMH endpoint can be offered. This is because Sciences Po's information systems department has been facing unprecedented demands in IT support and has subsequently been unable to allocate the required resources to set up the live version of Colectica Portal for the CDSP. As such, the CDSP was only able to launch the live version in early December 2021.

Nevertheless, despite these delays, the ethnic and migration studies data community has been able to complete a significant preparatory task related to the harvesting of the EMM QDB's metadata by the EQB in September 2021: documenting a full questionnaire using Colectica Questionnaires and Colectica Designer, for which an .xml file in DDI Lifecycle was produced and successfully read by CESSDA ERIC. Moreover, they have been able to leverage their ongoing collaboration with GESIS to explore the feasibility of including a new field to identify and distinguish formal collections (e.g. the EMM surveys to be harvested from the EMM QDB), as part of the CESSDA Metadata Model (CMM) (developed by the CESSDA Metadata Office (MDO))[49] and subsequently the EQB's metadata schema. To date, this new field was proposed and accepted by the CESSDA MDO, which suggests that the questionnaires for EMM surveys will also appear as a distinct collection on the EQB (and not just on the CDSP's Colectica Portal, as described in the previous section). Thus, in light of the aforementioned strides made, the ethnic and migration studies data community is still confident that before the SSHOC project concludes, it will be able to run a number of trials to assess the feasibility, including costs entailed, of harvesting the EMM QDB's metadata from the CDSP's Colectica Portal for the EQB.

[49] The User guide for the CMM, which includes detailed information about what it is and how it should be used, can be found on Zenodo: https://zenodo.org/record/3236194#.YbCDA_HML_Q [30 November 2021]

6. Conclusion: Evaluating the feasibility of setting up a collection of the EQB dedicated to EMM survey questionnaires and questions

The feasibility of setting up a collection within EQB of the questionnaires and questions (and their corresponding metadata) of the EMM surveys identified as part of the EMM Survey Registry work (also undertaken within Task 9.2) has been carefully explored. After the action steps of creating an EMM collection through the creation of an EMM QDB were set, realistic solutions to the anticipated challenges were identified.

Therefore, the position is that it is feasible to create an EMM collection for the EQB, particularly if the costs of scaling the present EMM QDB can be covered and the link between the EMM QDB and EQB is made to be durable and sustainable. At the same time, however, the coverage of the EMM QDB and EMM collection should be carefully re-considered and re-evaluated. Given that the EMM Survey Registry presently covers around 1,700 surveys, of which roughly 70% have a questionnaire available for re-use,[50] the number of questionnaires to be captured by the EMM QDB and EMM collection is significantly greater than what was initially anticipated. Coupled with the fact that most questionnaires (including its metadata) would be manually documented and translated using Colectica Questionnaires and Colectica Designer, which on average is currently estimated to take one FTE (full-time employee) half a month to complete one questionnaire, it would require significant human resources and time to cover questionnaires currently classified as eligible based on the existing inclusion criteria of the EMM QDB and EMM collection. Hence, either the inclusivity of the questionnaires to be documented needs to be reduced or the scale of the financial resources required needs to be carefully considered to sustain such an effort.

In the sections below, more details are provided about the estimated costs of creating a full collection of questionnaires of EMM surveys for the EMM QDB and EQB respectively. They also provide a discussion of the anticipated durability and sustainability of the link between the EMM QDB and EQB. As mentioned in the paragraph above, these insights are needed in order to properly contextualize the trials conducted to create an EMM collection via an EMM QDB and, in turn, determine the "right" parameters for scaling the present EMM QDB and subsequent EMM collection.

[50] The full list of surveys eligible for the EMM QDB and EMM collection (based on the current inclusion criteria to include any EMM survey for which the questionnaire is publicly available or available via a member of ETHMIGSURVEYDATA) can be found by consulting the EMM Survey Registry and filtering by "Publicly available" and "Available through a COST Action member" for filtering option, "8.16. AVAILABILITY OF THE SURVEY QUESTIONNAIRE FOR INDIVIDUAL RESEARCH USE."

6.1 Estimating the costs for the necessary technical solutions

To document the questionnaires and metadata for the EMM QDB in DDI Lifecycle, so that they can be successfully harvested by the EQB for the EMM collection, three different Colectica products will need to be secured. In **table 2** below, the estimated costs for purchasing all three products is provided, based on rough quotes procured from the Colectica website and users of these Colectica products.

Table 2: Estimated costs for Colectica products needed for setting up the EMM QDB and EMM collection of the EQB

Product	Purpose	Estimated costs
Colectica Questionnaires	Documenting the questionnaires and metadata of the EMM surveys	620 EUR per user for an annual subscription (at least three users for efficient progress needed = at least 1,860 EUR per year)
Colectica Designer		630 EUR per user for an annual subscription (only one required)
Colectica Portal	Displaying the documented questionnaires and metadata of the EMM surveys, on a public-facing interface	In the tens of thousands EUR for initial installation (already absorbed by the CDSP), plus annual maintenance in the thousands EUR (of which roughly 50-60% to be probably covered by the CDSP)[51]
NOTE: Generally, purchasing multiple products will result in reduced overall costs.		

The total costs of setting up the EMM QDB and EMM collection can thus be approximated, using the estimated costs from **table 2** above and based on the questionnaires to be actually covered by the EMM QDB and EMM collection. More specifically, there are three potential and realistic scenarios of questionnaire coverage for the EMM QDB and EMM collection (see scenarios 1-3 below) and their corresponding cost estimates have been calculated as follows:

Scenario 1 - all the surveys from the first thematic pilot, i.e. civic life and political integration: To date, roughly 400 questionnaires have been identified for inclusion. Assuming there would only be three users of Colectica Questionnaires, then the total estimated cost would be at least around 37,000 EUR (where the CDSP would probably cover around half of the Colectica Portal maintenance costs), as the work would take at least six years to complete.

Scenario 2 - all the surveys eligible for the first, second (i.e. stereotypes, prejudices, discrimination, intergroup relations/beliefs), or third (i.e. belonging and social identity, cultural and social norms) thematic pilot: The questionnaires eligible for the second or third thematic pilot have yet to be identified. However, assuming these two thematic pilots would cover around the same number of questionnaires as the first thematic pilot, to complete the work using the same working

[51] Unlike for Colectica Questionnaires and Colectica Designer, cost estimates for Colectica Portal are not provided on the Colectica website. As such, the estimated costs reported here are broad ballpark figures.

timeline as scenario 1 (i.e. six years) would cost at least around 72,000 EUR (i.e. nine users for Colectica Questionnaires per year (5,580 EUR) + one Colectica Designer license per year (630 EUR) + annual maintenance of Colectica Portal (assuming around 4,000-5,000 EUR, with CDSP covering around half)).

Scenario 3 - all the surveys with a questionnaire available for re-use: Of the 1,700 surveys currently displayed on the EMM Survey Registry, roughly 1,200 of them have a questionnaire available for re-use. If these 1,200 surveys had, for example, at least two questionnaires to be covered (which is likely to be an underestimation), then, using the six year working timeline (as was the case for scenarios 1-2), the total cost would be at least around 96,000 EUR (i.e. 17 users for Colectica Questionnaires per year (11,000 EUR) + one Colectica Designer license per year (630 EUR) + annual maintenance of Colectica Portal (assuming around 4,000-5,000 EUR, with CDSP covering around half)).

What these three scenarios, therefore, indicate is that securing the necessary Colectica products alone require a noteworthy financial commitment, even with the most restrictive questionnaire coverage of the EMM QDB and EMM collection (i.e. scenario 1). To this, the corresponding human resources in terms of person-months would need to be added, bearing in mind the high training costs and the skills required to contribute to this metadata input and curation. Furthermore, the three scenarios highlight the importance of selecting the appropriate inclusivity of questionnaires for the EMM QDB and EMM collection, in line with the available budget (which will also need to have funds earmarked for the human resources demands adapted to each scenario, as described next in section 6.2, Estimating the human resources needed).

6.2 Estimating the human resources needed

Based on the experiences and outcomes of the trials conducted via the first thematic pilot, it would be reasonable to budget two questionnaires per month for a single FTE (i.e. a full-time RA) in order to have all the question items and corresponding metadata to be properly documented, as well as all question items to be translated (assuming the FTE has the needed language skills and translation is actually needed). The number of questionnaires to be covered per FTE could potentially be increased to around 3-4 questionnaires if the semi-automatic extraction of structured text could be implemented for a significant number of questionnaires or if some of the questionnaires are already pre-edited in DDI Codebook through Nesstar or equivalent software used by CESSDA SP and could be imported. What this then means, in terms of the recruitment of FTEs, is that there are three important points for consideration: (1) the recruited FTEs need to be able to learn Colectica Questionnaires and Colectica Designer quickly and be able to execute the manual work of documenting questionnaires and metadata efficiently; (2) at least three FTEs need to be recruited per year, in order to allow for sufficient annual progress in documenting the questionnaires and corresponding metadata; and (3) the recruited FTEs need to cover collectively a wide-range of languages to ensure that translation work can be embedded as part of the questionnaire and metadata documentation processes. Particularly for the latter point, it is challenging to find RAs who have language skills for translating common EMM-spoken languages (e.g. Mandarin-English, Arab-English, Hungarian-English) and who can also perform the tasks related to quantitative survey research.

Beyond the actual work to document the questionnaires and metadata, including the translation of question items, there are also large human resource demands to manage the various workflows in place to create the EMM QDB and EMM collection respectively. First, there needs to be at least one dedicated FTE who can oversee and manage everyone who would be documenting questionnaires and the metadata, plus translating the question items as needed. Moreover, there needs to be a small

dedicated team who can manage the technical aspects of maintaining the Colectica Portal and the link between the EMM QDB with the EQB. Even with the limited number of questionnaires covered as part of the trials for this feasibility assessment, the ethnic and migration studies data community has seen the importance of having these two types of human resources in place. In fact, at a bare minimum, a correct implementation of the workflows requires two separate individuals at FTE (or close to FTE, e.g. 70-80%) to be able to manage the competing coordination tasks.

Therefore, what has been able to be confirmed, through this feasibility assessment, is that the work to construct the EMM QDB and the EMM collection requires a variety of human resources, all of which also come with their own financial costs. This then further bolsters the ethnic and migration studies data community's position that the number of questionnaires to be covered by the EMM QDB and EMM collection need to be determined with careful consideration of both the financial and human resources that can be realistically obtained.

6.3 Assessing the durability and sustainability of the EMM QDB and EQB link

At the time of completing this report, a live link is yet to be established between the EMM QDB (on the CDSP's Colectica Portal) and the EQB, as described in section 5.2, Harvesting the EMM QDB's questionnaires for the EQB. However, the essential preparatory work (for establishing the EMM QDB - EQB link) has been completed and validated by CESSDA ERIC (the entity now leading the development of the EQB) and both the CDSP and CESSDA ERIC are actively working together to test, as soon as possible, the harvesting of the EMM QDB's collection via the EQB. Therefore, despite the current state of the EMM QDB - EQB link, the EQB will, in a matter of short time, be able to harvest the EMM survey questionnaires, question items and their respective metadata from the EMM QDB and, subsequently, identify and validate the exact harvesting workflow to be implemented to set up the EMM collection. In fact, this is the main task that will be the focus of the Task 9.2 team in the four final months of the funding period of SSHOC (January-April 2022).

This then leads to the important question of whether the link, eventually set up, between the EMM QDB and EQB will be durable and sustainable. This is likely to be the case, given that both the CDSP and CESSDA ERIC have continuously articulated and communicated their intent to ensure that their respective services would be maintained in the long-term. CDSP has specifically budgeted and planned to run their Colectica Portal for the next ten years. CESSDA ERIC has specified that the EQB will be a live service where its workflows have been deliberately tested and designed to support routine curation and ingestion of questionnaires with a vocation of long-term sustainability.

Sustainability and durability are also affected by other aspects. On the one hand, the person-months allocated to this sub-task of Task 9.2 within SSHOC allowed for a pilot to conceive the service, find the appropriate technological tools, and establish the workflow protocols that would allow assessing whether creating an EMM QDB through the bottom-up impetus of the data communities and integrating it in the EQB was feasible. The pilot that has been reported on here suggests that it is, but it also suggests that a considerable amount of resources are necessary moving forward in order to sustain the human and budgetary resources needed to compile, organize, and input the hundreds of questionnaires that are relevant for an EMM QDB collection. Hence, this only seems sustainable and durable in the long-term if either the EMM QDB collection is absorbed by, or otherwise integrated in, existing infrastructures (e.g. CESSDA ERIC), or if a new ERIC focusing on research data and resources for

cross-disciplinary ethnic and migration studies were to be established. Both options will be explored in the months to come.

Furthermore, another sustainability and durability aspect that still needs to be assessed at a later stage is the possibility for existing and future data producers of EMM surveys to update and contribute to the EMM collection of the EQB. As the EMM QDB has not been designed to allow such data producers to add questionnaires for new EMM surveys on their own (as infrastructure support and access to the specialised software Colectica is needed to handle the inclusion of new questionnaires), then this updating of the EMM collection with new questionnaires needs to be addressed through other routes. One possibility would be that this contribution from the data producers could be done through EQB directly. However, this faces important limitations. With the emphasis of the EQB currently on establishing a stable and reliable curation and ingestion of questionnaires from CESSDA SPs, there is no functionality in place yet where data producers might be able to autonomously populate the EQB and, subsequently, the collection dedicated to EMM surveys. Moreover, even if such a functionality were to be provided, the EMM QDB collection provides much richer metadata particularly in relation to multilingual instruments as well as in relation to conceptual hierarchies and horizontal links that facilitate data discovery than EQB does. Therefore, a first alternative to having a feature within the EQB to contribute new questionnaires, is to have data producers of EMM surveys contribute new questionnaires by way of a CESSDA SP that could integrate the EMM QDB workflows as part of their standard questionnaire documentation workflows when EMM survey data are deposited. With this option though, CESSDA SPs would then need to establish workflows with the EQB to not only allow the EQB to recognize that these new questionnaires are intended for the EMM collection but also ensure that the new questionnaires are named, documented, and conceptually identified following the existing protocols for the EMM QDB. Finally, a second alternative would be (as discussed in the previous paragraph) to rely on an ethnic and migration studies infrastructure that would have the EMM QDB as one of its central services and tools.

7. References

Akdeniz, Esra. (2021). A report of the EQB Metadata Schema with a list of requirements to incorporate SPs' metadata into EQB (1.1). Zenodo. https://doi.org/10.5281/zenodo.4572440.

Colectica (2021a). Colectica. Colectica 6.2 Documentation. Retrieved November 12, 2021, from https://docs.colectica.com/.

Colectica (2021b). Colectica Designer. Colectica - Colectica Designer. Retrieved November 12, 2021, from https://www.colectica.com/software/designer/.

Krämer, Thomas, & Akdeniz, Esra. (2020). Euro Question Bank: Microservices and UI for fielded survey question search. 12th Annual European DDI User Conference (EDDI20). Zenodo. https://doi.org/10.5281/zenodo.4310903.

Mills, H. (2021). UCL Wiki. Archivist - CLOSER. Retrieved November 12, 2021, from https://wiki.ucl.ac.uk/display/CLOS/Archivist.

Saji, Ami & Laura Morales. (2020). D9.4 Database with the metadata of surveys to EMMs across Europe (v1.0). Zenodo. https://doi.org/10.5281/zenodo.4558307

SERISS (2019). Why do social researchers need the QVDB? Retrieved November 23, 2021, from https://seriss.eu/wp-content/uploads/2019/08/SERISS-Info-sheet_QVDB.pdf.

List of Images

List of Tables

Appendix

A: Preliminary analysis of EMM surveys to be included in the first thematic pilot (excerpt from a memo circulated in 01.2020)

As of January 2020, **final metadata files from all the pilot participating countries have been received, except for France**. From these files, **254 surveys** (n=152 for national, n=102 for subnational) **have been identified as covering the topics 1.13.20, 1.13.24 and/or 1.13.27**. The following table provides a more detailed breakdown of these 254 surveys, in relation to the availability of the questionnaire and the quality ranking.

	A. Surveys covering topics 1.13.20 and/or 1.13.24 ***(by availability of the questionnaire and quality rating)***			
		All	**National**	**Subnational**
Questionnaire publicly available *["1.Publicly available" for 8.16]*	**All**	**165**	**106**	**59**
	Rating of 6 or greater	**121**	89	32
	Rating of 5 or less	**44**	17	27
Questionnaire available by request or through COST Action member *["2.Available through a COST Action member" or "3.Available by request" for 8.16]*	**All**	**39**	**26**	**13**
	Rating of 6 or greater	**37**	25	12
	Rating of 5 or less	**2**	1	1

Moreover, **41 different languages** have been identified in the **254 surveys by examining 8.20 ("Language(s) in which the survey questionnaire for individual research use is available") of the files** [*NOTE: The language count is a rough estimate and is also likely to be an underestimation. Not all the templates had variable 8.20 filled out and not all the templates properly listed the languages (e.g. encoding responses like, "All languages spoken by the migrants"). Moreover, there is another variable, 7.7, that looks at the questionnaire languages. The languages listed in 8.20 and 7.7 are not always the same, and in many cases there were fewer languages listed in 8.20 than in 7.7 (e.g. the data producers only made the English version of the questionnaire available and have kept the questionnaires in the migrant languages for their personal records only.*]. National surveys covered 39 of the 41 languages, whereas the subnational surveys covered 19 of the 41 languages. In taking into account the availability of the questionnaire and the quality rating, tables B-D illustrate the detailed breakdown:

B. Number of languages identified for national and subnational surveys *(by availability of the questionnaire and quality rating)*

Category	Count
All	**41**
Publicly available	35
Available by request or through a COST Action member	23
Publicly available with rating of 6 or greater	32
Publicly available with rating of 5 or less	15
Available by request or through a COST Action member with rating of 6 or greater	23
Available by request or through a COST Action member with rating of 5 or less	3

What the tables above illustrate is that the number of languages to handle might not change enormously (e.g. from 31 to 22 or 29 if we establish some criterion of ease of availability or quality). In terms of the types of languages identified, they include not only European languages but also a wide range of migrant languages. This diversity in languages is important to note, as it has implications for how translations between English and the other language would be validated. Specifically, it would need to be determined as to how to find individuals with the linguistic capabilities to validate translations between English and languages such as: Arabic, Bashkir, Bengali, Dari, Farsi, Gujarati, Pashto, Punjabi, Somali, Tagalog, Tamil, Tigrinya, and Urdu.

C. Number of languages identified for national level surveys ***(by availability of the questionnaire and quality rating)***	
Category	**Count**
All	**39**
Publicly available	31
Available by request or through a COST Action member	22
Publicly available with rating of 6 or greater	29
Publicly available with rating of 5 or less	8
Available by request or through a COST Action member with rating of 6 or greater	22
Available by request or through a COST Action member with rating of 5 or less	1

D. Number of languages identified for subnational level surveys ***(by availability of the questionnaire and quality rating)***	
Category	**Count**
All	**19**
Publicly available	18
Available by request or through a COST Action member	5
Publicly available with rating of 6 or greater	12
Publicly available with rating of 5 or less	11
Available by request or through a COST Action member with rating of 6 or greater	4
Available by request or through a COST Action member with rating of 5 or less	2

B: EMM QDB workflow: retrieving questionnaires (copy of working draft last updated on 30.09.2021)

Introduction

This document outlines the workflow for retrieving questionnaires for the three EMM QDB pilots, which will be co-developed and co-implemented by COST Action 16111 - ETHMIGSURVEYDATA; the Task 9.2 team of the H2020-funded project, SSHOC (Social Sciences and Open Cloud); and an ANR Flash-funded project, FAIRETHMIGQUANT. The pilots will be building on the EMM Survey Registry work that has been undertaken by ETHMIGSURVEYDATA and the Task 9.2 team, as it will use questionnaires from surveys that have been or will be included in the EMM Survey Registry. The main objective of the pilots is to test the feasibility of setting up a full scale and EMM-dedicated collection for the CESSDA ERIC-led Euro Question Bank (EQB).

The pilots will specifically include questionnaires from ten geographically and linguistically diverse European countries that also produce different types of quantitative surveys on ethnic and migrant minorities' (EMMs') integration: Croatia, France, Germany, Hungary, Italy, the Netherlands, Norway, Spain, Switzerland, and the UK. It will also be limited to three thematic topics that will each be implemented as its own pilot in the order they appear below:

- **Pilot 1 - Civic life and political integration:** political inclusion and participation, political attitudes, social cohesion, civic engagement, social networks and ties, transnational patterns, diasporas (variables 1.13.20, 1.13.24, and 1.13.27 of the survey metadata template used to develop the EMM Survey Registry)
- **Pilot 2 - Stereotypes, prejudices, discrimination, intergroup relations/beliefs:** discrimination, racism and/or xenophobia, interethnic contact and conflict, public attitudes about migration and migrants (variables 1.13.5, 1.13.14, and 1.13.21)
- **Pilot 3 - Belonging and social identity, cultural and social norms**: identity (ethnic, national, racial, religious) and belonging, gender relations, gender identity, sexuality, social attitudes, social values (variables 1.13.8, 1.13.12, and 1.13.20)

Step 1: Collecting the questionnaires

The central team at Sciences Po (Laura Morales, Meredith Winn, Ami Saji, RAs/vacataires/interns employed by SSHOC or FAIRETHMIGQUANT) will be responsible for collecting the questionnaires. The questionnaires should be collected using the following procedure:

1. Using the EMM Survey Registry and its filtering options, **identify surveys (i.e. unique records on the EMM Survey Registry) that meet ALL of following criteria**:
   a. Has been conducted in one of the 10 pilot countries (**variable 1.0**),
   b. Covers AT LEAST ONE of the variables identified within a thematic topic (**pilot 1 = variables 1.13.20, 1.13.24, 1.13.27 | pilot 2 = variables 1.13.5, 1.13.14, 1.13.21 | pilot 3 = 1.13.8, 1.13.12, and 1.13.20**) AND

   c. Has a questionnaire that is indicated as either "Publicly Available"—i.e. is available online and can be downloaded (even if an account must be created to do so or—or Available by through a COST Action member" (**variable 8.16**).

2. For each survey that **fully meets the criteria** above, **add it to the Google Drive-based tracker**, ***EMM QDB Pilots Questionnaire Tracker***,[52] and **fill out information** for all of the **columns marked with a single asterisk (the first cell of these columns is indicated in dark blue)**. The relevant columns will collect information on the following: the country in which the survey was conducted (variable 1.0); whether the survey is part of a cross-country survey program (variable group 2.c); the survey name in English and in the native language (variables 1.3 and 1.4); the relevant thematic topic variables (variable group 1.13); the availability of the dataset (variable 8.1), including details on how to obtain it (variable 8.2); and the availability of the questionnaire (variable 8.16), including details on how to obtain it (variable 8.17). The variable numbers have been indicated in the first row of the column in parentheticals following the asterisk.
   [NOTE: For surveys that are part of a cross-country and/or multi-wave survey program, steps 3-5 listed below may need to be handled differently. Please read and apply the specific instructions provided on the INSTRUCTIONS tab of the EMM QDB Pilots Questionnaire Tracker.]
3. **Collect the questionnaires for the surveys** that have been added to the *EMM QDB Pilots Questionnaire Tracker*. Point a below should be consulted whenever a questionnaire is publicly available and point b below should be consulted whenever a questionnaire is available through request from a COST Action member.
   a. **Questionnaire is publicly available:**
      i. If the publicly available questionnaire has been **made available as a single file** (e.g. There is 1 PDF document that has the English and Norwegian versions of the questionnaire), download it and then **upload it to the *EMM QDB workspace in MyCore***,[53] a document storing and sharing platform that is operated by the CNRS. The questionnaire should be deposited into the folder for the relevant thematic pilot (e.g. **1. Pilot 1** for a questionnaire qualifying for the first thematic pilot). The uploaded questionnaire should also be **named using the following naming convention**: Survey acronym (survey year).documenttype [*EXAMPLE: LOCALMULTIDEM(2004).pdf or SCIP(2010).docx*].
         - *NOTE 1: Information on the survey acronym can be found in variable 2.2 for cross-country and/or repeated cross-sectional/longitudinal surveys and in variable 1.2 for all other types of surveys.*
         - *NOTE 2: If there is no survey acronym, please use the survey name in English (variable 2.3 for cross-country and/or repeated cross-sectional/longitudinal surveys and in variable 1.3 for all other types of surveys) instead [EXAMPLE: Sustainability of Minority Return in Croatia (2006).pdf]*
         - *NOTE 3: Information on survey year can be found in variables 1.10-1.11, except for cross-country surveys. For such surveys, please use the start date for the cross-country survey as a whole, which is information typically found on official documentation about the survey.*
      ii. If the publicly available questionnaire has been **made available as multiple files** (e.g. There are 2 PDF documents for the questionnaire, one for the English version and one for the Norwegian version), open up the ***EMM QDB workspace in MyCore*** and go to the folder for the relevant thematic pilot. Create a new folder for the questionnaire files **using the**

[52] The actual link to the *EMM QDB Pilots Questionnaire Tracker* is provided in the actual workflow document. For security reasons, the link has been removed here.
[53] The link to the *EMM QDB workspace in MyCore* is provided in the actual workflow document. For security reasons, the link has been removed here.

following naming convention: Survey acronym (survey year) [*EXAMPLE: ICS (2012)*]. Download all of the questionnaire files and then add them to the new folder. Please refer to the *Naming Convention Key*[54] for naming files in folders with multiple files.

- *NOTE 1: Information on the survey acronym can be found in variable 2.2 for cross-country and/or repeated cross-sectional/longitudinal surveys and in variable 1.2 for all other types of surveys.*
- *NOTE 2: If there is no survey acronym, please use the survey name in English (variable 2.3 for cross-country and/or repeated cross-sectional/longitudinal surveys and in variable 1.3 for all other types of surveys) instead [EXAMPLE: Sustainability of Minority Return in Croatia (2006).pdf]*
- *Information on survey year can be found in variables 1.10-1.11, except for cross-country surveys. For such surveys, please use the start date for the cross-country survey as a whole, which is information typically found on official documentation about the survey.*

b. **Questionnaire is available through request from a COST Action member:** For any survey for which the **questionnaire needs to be requested from a COST Action member**, please double check that this has been properly documented on the *EMM QDB Pilots Questionnaire Tracker*. Please then alert Meredith, Ami and/or Laura about this survey, so they can provide assistance in drafting an email that will allow the questionnaire to be properly requested from the relevant COST Action member. Once the questionnaire has been retrieved from the relevant COST Action member, follow the instructions in point a above about how to add a questionnaire to MyCore.

4. **After a questionnaire has successfully been added to MyCore**, **fill out all the cells corresponding to a double asterisked column (double asterisked columns are also headed with dark blue)** on the *EMM QDB Pilots Questionnaire Tracker*. The relevant columns will collect information on the following: who retrieved and uploaded the questionnaire to MyCore, when the questionnaire was retrieved and uploaded to MyCore, how the questionnaire has been named in MyCore, where in MyCore the questionnaire was uploaded, whether the questionnaire is available in English, and what other languages the questionnaire has been made available.
   - *NOTE: For a survey that is part of a cross-country and/or multi-wave survey program AND the questionnaires from at least more than 1 country/wave have been provided through a single file, please fill out the language-related cells using the questionnaire languages for the country/wave you are working on ONLY.*
5. **Double check how the language(s) of the retrieved questionnaire has/have been coded for variables 8.20** (language availability of the questionnaire available for research purposes) **and variable 7.7** (language availability of the questionnaire available for fieldwork) for the corresponding survey record on the EMM Survey Registry. If the languages of the retrieved questionnaire and variables 8.20 and 7.7 match, no further action is needed. However, if the languages of the retrieved questionnaire and variables 8.20 and 7.7 do not match, add a comment about this to the relevant cell on the *EMM QDB Pilots Questionnaire Tracker* and then notify Meredith, Ami and/or Laura. Meredith, Ami and and/or Laura will then provide guidance on how to liaise with the COST Action member who compiled the survey record on the registry to resolve this discrepancy, as well as when and how the survey record on the EMM Survey Registry and the data file being used for analysis need to corrected for the questionnaire-in-question.

[54] The actual link to the *Naming Convention Key* is provided in the actual workflow document. For security reasons, the link has been removed here.

C: EMM QDB workflow: documenting questionnaires using Colectica (copy of working draft last updated on 09.12.2021)

Introduction

This document outlines the **workflow for documenting an EMM Question Data Bank (QDB)-eligible questionnaire (and its metadata) using Colectica**. Currently, the EMM QDB is in its pilot stage, and this document is being developed in conjunction with the first thematic pilot: civic life and political integration (in other words, the steps outlined below are specific to this pilot only). However, the goal is that this workflow can be replicated on a larger scale for the full EMM QDB.

Step 1: Accessing the questionnaire

Each questionnaire must be documented using Colectica Questionnaires. In order to do this, a copy of the questionnaire must be retrieved.

1. Retrieve the questionnaire from the *EMM QDB workspace in MyCore*. To retrieve it, go to the folder: 1.Pilot 1. Questionnaires are named according to the name of the survey. If there are multiple questionnaires for the same survey (for instance, multiple language versions), these will be stored in a folder. Be sure to verify that you download the correct language version and (where it applies) year.
2. Download the questionnaire onto your machine. Be sure to save it to a place where it is easily accessible.

Step 2: Documenting the questionnaire

1. Update the *Colectica Questionnaires Tracker*[55] with information on the questionnaire you are inputting.
2. Use Colectica Questionnaires to document the entire questionnaire according to the Guidelines.[56]
3. Export the questionnaire as a PDF and check it for any obvious errors.
4. Export the questionnaire to DDI 3.2. Be sure the file name includes the survey name, necessary identifying issue, and final_unmapped.
5. Publish the questionnaire to the pre-production portal by clicking 'Synchronize'.
6. When the questionnaire has been fully input, indicate this in the aforementioned *Colectica Questionnaires Tracker* and move on to step 3 below.

Step 3: Conceptual mapping

1. Download the questionnaire from the remote repository and open on Colectica Designer.

[55] The actual link to the *Colectica Questionnaires Tracker* is provided in the actual workflow document. For security reasons, the link has been removed here.
[56] The actual link to the *Guidelines* is provided in the actual workflow document. For security reasons, the link has been removed here.

2. Also open the *Concept spreadsheet*[57]for the first thematic pilot for reference. You will also need to have the *Question Label Tracker*[58] open.
3. Beginning with the Question Constructs, open each question item to check for relevance to the defined concepts for the first thematic pilot (NOTE: All concepts have been created and entered for the first thematic pilot. If new concepts are added or the hierarchy is altered, follow the Colectica Designer Guide to incorporate these changes). When you click on the question item, you will be able to read the question text. This is also a moment to quickly check for potential typos in the question name, text, or instructions.

****For question items falling within the first thematic pilot's thematic fields****

4. Copy the Question Name and paste it into the Question Name column in the aforementioned *Question Label Tracker.*
5. Navigate to the Conceptual tab and add the relevant concept or concepts to the question item. Be sure to select an existing concept from the list (using the magnifying glass). DO NOT create a new concept (NOTE: If you accidentally create a new concept, be sure that it is blank and remove it from the question item. It will be deleted when the local database is cleaned).

****Steps 6-11 only apply for questionnaires that have multiple language versions, are fielded in multiple cities, or that are longitudinal. If it is a stand-alone questionnaire, skip to step 11****

6. In the Represented Variable box, type concept name followed by (SurveyAcronym-Extrainfo) in accordance with the naming and labeling convention (NOTE: You could also copy the concept name by clicking on the concept which redirects you to the concept metadata. Here, you copy the concept name from the label box. Then, use the tabs at the top (you may need to scroll horizontally) to return to the question item).
7. Click on the Represented Variable.
8. Paste the name of the question into the Name box. Then, remove the language information. This creates a name consistent with the naming and labeling convention.
9. Copy the concept name from the label box and paste it into the Concept column of the aforementioned *Question Label Tracker* (NOTE: The order of copying and pasting may be altered slightly if you copy the concept via the process described in the above footnote).
10. Now, you need to add the Conceptual Variable. There are two processes depending on if the conceptual variable has already been created (as part of a related questionnaire).
    a. *If the Conceptual Variable Must Be Created* (NOTE: There should only exist 1 conceptual variable for each concept within a single study (longitudinal, cross-national, or multilingual). Conceptual variables for the same concept should be unique to the study): Click the + icon to create a new conceptual variable. In the label box, type or paste the concept name followed by the study acronym in parentheses (seeing naming and labeling conventions). Attach the corresponding concept to the conceptual variable
    b. *If the Conceptual Variable has already been created*: Use the magnifying glass to find and select the appropriate Conceptual Variable.
11. Repeat these steps for all Question Items in the Question Constructs.
12. Once Question Constructs have been completed, do the same exercise with sequences. Some of the question items may have already appeared under Constructs, but you should verify that all have been covered.

[57] The actual link to the *Concept spreadsheet* is provided in the actual workflow document. For security reasons, the link has been removed here.
[58] The actual link to the *Question Label Tracker* is provided in the actual workflow document. For security reasons, the link has been removed here.

13. Once all question items have been mapped, save the results and publish them to the repository.

Step 4: Labeling Question Items

Question Labels use conceptual information. Therefore, they are added after the conceptual mapping has been completed using the aforementioned *Question Label Tracker*.

1. Open Colectica Questionnaires. Open the desired questionnaire **using Open from Repository**
2. For each question item in the aforementioned *Question Label Tracker*, add the label following the accepted convention:
    a. [Concept] (StudyAcronym-Year-Location-Language) (e.g. LMD-Lyon-FRA) (NOTE: Study acronym and year location only if applicable).
3. Indicate in the *Question Label Tracker* when you have added the question label.
4. Once all question labels have been updated save the results. First, hit the save icon. Second, export the metadata as DDI (indicate the survey and date in the title and indicate final_final). Third, publish to the online repository by clicking on Synchronize.
    a. Be sure that the items are published before closing the program.
5. Update the aforementioned *Colectica Questionnaires Tracker*.