Title	**Recommendations for, and requirements of, a CESSDA Harmonisation Infrastructure (D9.4)**
Work Package	WP9
Authors	Markus Quandt, compiled from team contributions
Dissemination Level	PU (Public)

Summary/abstract

This document is a brief compilation of recommendations for a CESSDA harmonisation infrastructure, in which a question database plays an important supportive role. Requirements for the implementation of this infrastructure are listed. These requirements include most of the more advanced technical and organisational goals that the CESSDA-PPP has formulated for the future CESSDA infrastructure. Thus, a full-fledged implementation of the harmonisation infrastructure cannot be pursued independently from developing the basic CESSDA infrastructure itself. Both must be implemented in well-planned steps. In that process, the harmonisation infrastructure can at any stage of the development be regarded as a show case for the capabilities of the underlying CESSDA infrastructure.

1 Introduction....2
2 The Data Harmonisation Platform (3CDB)....3
2.1 Core Recommendations....3
2.2 Technical Requirements of 3CDB....6
2.3 Best Practice Expectations for the 3CDB....7
2.4 Future Resource Needs of the 3CDB....7
3 The Question Data Base (QDB)....8
3.1 Core Recommendations....8
3.2 Technical Requirements of QDB....11
3.3 Best Practice Expectations for the QDB....11
3.4 Future Resource Needs of the QDB....11
4 Collaboration Options....12
5 Policy Recommendations....12
6 References....13

1 Introduction

Social science research and policy making within EU and beyond are increasingly based on comparative analysis of survey data (Atkinson et al., 2003). A host of methodological work is focused on evaluating the validity of such comparisons across data from different sources. In striking contrast, much less attention is being paid to the practical task of making existing survey datasets more comparable.

Survey programmes such as the European Social Survey or the EU Labour Force Survey, which design their surveys for comparability from the outset, are usually in a rather comfortable position in that they require little ex-post harmonisation, except for variables such as educational degree that by their nature depend on national institutions and thus cannot be collected in a fully standardized manner. But beyond that, vast amounts of data not originally collected with comparison and comparability in mind can be highly relevant for comparative research. This is true for almost all data collected by national statistical offices; it also holds for much academic research, such as electoral surveys; more generally, it holds for all data sets that address topics having some necessary similarities across countries or points in time. With such data becoming accessible and searchable through European or even larger infrastructures (cf. the DataVerse project, King, 2007), the demand for ex post-harmonisation will increase with certainty.

The term “data harmonisation” here refers to the process of transforming data from different sources into standard measures that facilitate undertaking research involving comparisons over the sources. The different data sources often stand for discrete realisations of dimensions such as time and space, comparisons being made between points in time, or between countries. Typically, data harmonisation involves the creation of “conversion keys” where the values of one or several source variable(s) are transformed into new values of a standardised target variable. When performing such data manipulations to improve comparability, researchers often fall back on ad hoc considerations, using a diversity of tools which they select by the almost random criteria of familiarity and availability, and usually creating no or little documentation.

Harmonisation work in itself is usually not regarded to be worthy of a self-contained publication and hence it is not accessible in peer-reviewed articles or books. Currently, a lack of incentives for researchers results in a lack of supply of proper harmonisation work. The rare exceptions are some dispersed semi-official documents from bodies such as the OECD or EUROSTAT on harmonising certain official statistics, guidelines on harmonisation of demographic variables worked out internally by ESOMAR and the ESS and ISSP methodological groups, or a few websites of specialized researchers. Publication channels and infrastructures that facilitate the creation, documentation and dissemination of harmonized variables, do not exist at present or are at best in an incipient phase (e.g., GE*DE, CCESD-IS).

In response to this situation, CESSDA has set out to develop a platform which would provide a common framework for documenting, distributing and applying the results of harmonisation efforts (CESSDA, 2008-2009; CESSDA PPP - Work Package 9, 2008). More specifically, two specific technical and organizational facilities were to be drafted within Work Package 9 of the CESSDA-PPP:

1. A database designed to make existing routines and knowledge on harmonisation of variables more widely available, named Concepts, Conversion, Classification Database (3CDB), and
2. A question database (QDB) to offer access to question texts and question metadata, and/or to survey results connected to specific questions.

Both facilities are to be designed such that they can act as common resources in the CESSDA infrastructure, giving access to the distributed data and metadata holdings of CESSDA members, and supporting users throughout Europe and beyond.

2 The Data Harmonisation Platform (3CDB)

2.1 Core Recommendations

The solution proposed for fostering harmonisation work in the CESSDA-ERIC is a central online platform, which either directly stores or connects all components that are relevant for performing data harmonisation work, in particular ex-post data harmonisation. Core components are concepts, classifications, and conversion routines, to be collected in a common database, therefore the name of "3CDB" was chosen. The purpose of the 3CDB (synonyms: CCCDB, CHARMCATS[1]) is to support the creation, storage, access to, and distribution of harmonised variables for comparative (cross-cultural, longitudinal, multi-group, multi-level) social research. A full description of the 3CDB, comprising the basic workflow concepts and the interface design of the harmonisation workbench part of the application as well as a full data model, is given in the report of Task 9.4/Deliverable 9.2 (Bourmpos et al., 2009) and will not be repeated here. A proof-of-concept application (CHARMCATS) with the user interface and some basic functionality has been developed and will be submitted with the final collection of project materials.

The goal of the 3CDB platform is to create a central database for harmonisation routines that offers the following services:

- to create, store and publish descriptions of Classifications, Scales, and Indices (CSI);
- to create, store and publish harmonisation (conversion) routines (CR);
- to enrich all elements with detailed and transparent documentation;
- to connect all of the above to (metadata on) variables, questions and data files, yielding a complete harmonisation project (HP);
- to derive new CSI and CR from existing CSIs and/or related CR;
- to assist in applying the CR to the data (data manipulation);
- to publish harmonisation projects and make them citable in analogy to printed research publications, through the use of persistent identifiers.

With methodological documentation of classifications and other measurement instruments being natural parts of any harmonisation project, 3CDB will automatically become a database of such methodological information, too. However, this is not a priority goal.

[1] Cessda HARMonisation of CATegories and Scales

A basic tenet of the harmonisation workbench feature of 3CDB is a 3-step working approach. The first step – conceptual – comprises theoretical and conceptual definitions, the second step – operationalisation – concerns adaptation to contextual specificities of the different data sources, the third step – data coding – concerns adaptation to the peculiarities of the actual data files. It is expected that this will increase the scientific value of the harmonization work: instead of having only unstructured documentation of source to target variables and re-coding, which would simply help replicating a particular conversion syntax, all potential sources of bias and equivalence (conceptual, operational and data specific) can now be systematically evaluated and stored. Furthermore, different authors' contributions to each different step can now be made visible. This is hoped to serve as an additional incentive for 'granular' contributions, even when the individual contribution is limited to publishing single components of a harmonisation project. The application will support the 3-step workflow through its graphical interface.

As a final stage of the implementation, we propose an online system for the 3CDB that borrows some of the basic principles of collaborative publication work from the model of Wikis. Technically, it is envisioned that the final system will be a web platform allowing collaborative building of entries via a public access interface. Collaboration can occur in different ways: (1) multiple users may work on the very same harmonisation project *before* it is published, thus allowing use by research groups. (2) Different users may build on previously published harmonisation projects, by modifying elements of these – for example, where they disagree with the original creators for scientific reasons –, or by adapting the conversion routine of a harmonisation project to a new set of source data. Besides active collaboration in contributing to new entries, the system will also provide search and download access for non-contributing users.

Differences to common Wiki approaches arise from the well-established norms around scientific publications: Contributions are clearly to be assigned to their originators to maintain the incentives for earning credit and responsibility. Also, contributions must be frozen at the very moment of final publication to allow for scientific replication. 'Edits' to a public entry always lead to the creation of a new instance of the original entry; thus, editing users create derivations and not modifications. This does not preclude adding comments to entries, by way of an open discussion forum. Derived entries are always automatically marked as such, and contain references to the original contribution. Further, published harmonisation projects will be assigned persistent identifiers that enable their citation in publications and other harmonisation projects.

WP 9 considered three contribution approaches for the harmonisation infrastructure, spanning from a very "centralized" production approach (where every input in the database will be controlled and approved by CESSDA) to a totally "open approach" (where anyone can contribute to the database). WP 9 finally recommended adopting a "controlled contribution approach" which falls somewhere between these two poles (Quandt et al., 2008).

Under the controlled contribution approach, there will be two main types of researchers using the system: (1) contributors using the platform as a workbench for creating harmonisation routines, and (2) more passive users, exploring, finding and

perhaps evaluating and re-using existing materials. Under the first view, a desktop prototype is being developed as a proof-of-concept application under the name of CHARMCATS.

Given the nature of harmonisation work – which connects data across multiple sources, not only sources of data collection, but also sources of data dissemination –, it is crucial that the online platform works as a single point of access to *all* of the distributed data holdings of CESSDA members. Therefore, its interface should be linked into the CESSDA portal. It is, however, not technically required that the platform software and database run on the same servers as the CESSDA portal, or that they are maintained by the same CESSDA member organisation.

The online harmonisation platform can be implemented in sequential stages.

- The first stage would simply be a database of well-documented conversion routines, which would be linked into the CESSDA infrastructure in a passive manner. 'Passive' means that it would be reading metadata and data from other CESSDA (members') servers, but not write back anything to those servers. This first stage is very much like the current proof-of-concept application, plus online searches across CESSDA holdings through the user interface of the harmonisation platform. However, the user management system must be in place in the first stage already, allowing 'contributing' and 'reading' users the access required to contribute content and comments, by which they build the holdings of the 3CDB. Referencing objects (data sets, variables, questions) in the CESSDA infrastructure (instead of duplicating them into the 3CDB database itself) can optionally be made a feature of the first phase, or moved to a subsequent stage, depending on when persistent identifiers are available throughout CESSDA resources.
- Subsequent stages would feed back (write to other servers, or expose for searches through standardised interfaces) material into the CESSDA infrastructure. There can be different exposable types of material. Provision of these again can be implemented in separates stages:
    - Coding routines for harmonised variables, or harmonised variables as such in form of partial data sets.
    - Comparability metadata, such as information on equivalent indicators used in different data sources.
    - Any new study, variable, or question level metadata entered by users of 3CDB that may be of interest to users of the CESSDA infrastructure beyond the 3CDB.
    - 3CDB is also one possible source for knowledge products such as international standard codes (ISCO for job classifications, ISO3166 for country codes, etc.).

Because the implementation of the 3CDB heavily depends on the prior or parallel implementation of the extended CESSDA infrastructure, it is not possible to propose a timeline from the perspective of WP9 alone. One important consideration is that building a substantial community of contributors and users may take at least two years after the initial public release. However, the attractiveness of the online platform will increase progressively with the number of available entries and contributions. A high number of contributing 3CDB users would underline CESSDA's potential to support

cutting edge research directly, which would be a crucial addition to the currently more basic services of the CESSDA archives.

There are internal strategic benefits for CESSDA as well: Among other things, 3CDB is a tool to organise the workflow of creating comparability information on existing data sets. If one takes into account that all of the present data holdings are document in DDI 2 format, but that the output of 3CDB must be DDI 3 – because no other standard able to handle comparability information exists – it is clear that 3CDB is a tool that helps in the selective translation of DDI 2 studies into DDI 3 information. Specifically, 3CDB by its researcher targeted approach addresses those areas (of those studies) where the labour intensive intellectual addition of DDI 3 information is worthwhile, given a particular research purpose. However, CESSDA still must answer the question of how the DDI 2 and DDI 3 information can be linked to each other, yet without inventing an additional DDI dialect with undesirable longevity.

2.2 Technical Requirements of 3CDB

Both of the online systems, 3CDB and QDB, make assumptions about the infrastructure they are imbedded in (see section 3 below for more information on the recommended infrastructure). Implicitly, these are assumptions about contributions that CESSDA members make towards that infrastructure, in the form of making their data provision systems interoperable with the infrastructure. Overall, 3CDB is more demanding than QDB, but even 3CDB is not posing demands that had not already been considered when the CESSDA-PPP and its work packages were devised.

Requirements for a first stage implementation of 3CDB (only read access to other CESSDA materials) could be:

- Access to detailed variable level metadata, including full question text metadata and minimal aggregate statistics such as frequency distributions and/or descriptive statistics for distributions, ideally, also dynamic access to the raw data of the relevant variables. Access to detailed question (and related study) metadata can be provided through a QDB or a similar system.
- Online dissemination systems (such as NESSTAR servers) for data of relevance to harmonisation work, and the QDB, must respond to standardised search and data provision queries. The content provided in response to such queries must follow at least the DDI 2 standard, however preferably DDI 3, so that 3CDB could import comparability and linking information that is available elsewhere. A protocol must be developed that encapsulates DDI 2/3 information in a format viable for live online requests and machine-to-machine communication.
- 3CDB assumes transparent handling of user authentication to distributed CESSDA resources, to be able to handle searches over these resources and data retrieval from them. I.e., full single-sign-on must be in place. It is desirable, but not strictly necessary, that registration as a new user of 3CDB – which includes an approval step through the administration team of 3CDB – can access existing contact information and credentials of the user elsewhere in the CESSDA infrastructure.
- If referencing is implemented at this stage: A reliable PID system. This includes real persistence of all metadata and data ever published, because any such data can be the target of references made in the 3CDB. In other words,

not only must the identifiers of data remain intact, but also the identified data itself must remain accessible to live requests. This explicitly includes data for which newer versions or revisions have become available.

The second implementation stage of 3CDB will need:

- Certainly: the ability of other CESSDA resources to digest extensive comparability information on the feedback route; therefore, DDI 3 compatibility.
- Possibly: a data manipulation engine in NESSTAR or a similar data dissemination system, which executes harmonisation routines dynamically against data held in distributed/federated servers. This data manipulation engine could reside on the CESSDA portal, which per default has access to all distributed data resources. It should be able to write the target variable of a harmonisation project to a new column of data, which users then can merge seamlessly into a virtual data set for online analysis or into the user's local data file.

A list of DDI 3 elements used internally by 3CDB is already available in the D9.2 report. This will have to be extended to cover the elements used in metadata ingest and metadata exposure of 3CDB.

2.3 Best Practice Expectations for the 3CDB

Best practices of data harmonisation work are inherently incorporated in the interface of the 3CDB application. The interface is designed to support a structured workflow and will distinguish between mandatory, recommended, and optional elements of documentation of a harmonisation project. Nevertheless, a best practice guide describing the ideal workflow and the methodological guidelines behind that should be made available.

The concept for the 3CDB explicitly allows for the situation that metadata retrieved from online sources have to be supplemented manually with lacking information in the 3CDB application itself. Thus, the 3CDB is not technically restrained by possible gaps in the material provided by remote CESSDA resources. However, it is obvious that the system is the more attractive to users, the less manual metadata entry they have to perform. Therefore, harmonisation work in general and 3CDB in general would benefit massively from CESSDA giving itself strict rules for the completeness of published metadata.

2.4 Future Resource Needs of the 3CDB

3CDB and QDB are consciously designed to be user-driven in terms of collecting content with added value beyond the basic CESSDA data and metadata holdings. With regard to their integration into the larger CESSDA infrastructure, we expect 3CDB and QDB to rely mostly on automatic, machine-based data access and exchange. Thus, the most relevant resource block is required for initial development and implementation. It is reasonable to expect at least two FTEs for two years for the initial development of the 3CDB in its first online implementation stage.

After that, we expect that technical maintenance and substantive administration of both systems do not require fulltime positions, but can be distributed across permanent work groups installed in the CESSDA network. Only the organisations

employing convenors of the work groups might need special funding. As any such funding models should follow future CESSDA rules for similar situations, the funding structure as such is not further discussed here. Some indication of the work groups can however already be given. The present proposal postulates three areas that require intervention by staff with substantive expertise in harmonisation (cf. Interim Report of T9.3). The first is the area of providing an initial set of harmonisation projects into the database, in order to make available at least some content when the platform first goes public. This should be continuously extended with harmonisation projects that are of central relevance to CESSDA work and are therefore provided or approved by competent CESSDA staff. The second area is that of approving the credentials of users who want to register as contributors to the system, and is therefore a continuous task from the point of the public release. The third area is that of loosely monitoring the contents provided, to remove 'garbage' entries, and to deal with obvious IP or other violations, and will also require continuous staff attention. While the T9.3 interim report defines distinct responsibilities in these areas for up to five teams (user administration, editorial monitoring, content harvesting, expert panel for quality approval, technical maintenance), it is obvious that these responsibilities can be grouped together, for the case that appropriate staff resources are scarce. It seems most efficient to assign responsibilities to a limited number of CESSDA members according to already available competencies.

3 The Question Data Base (QDB)

3.1 Core Recommendations

Even more than the 3CDB, the QDB can only be conceived of as an online system with a central access interface. The core objective for a QDB is to provide as much coverage of questions ever used in surveys as possible. Currently, the most effective way to realise this is to give access to all CESSDA holdings that are already documented at the variable level, because this usually includes full question texts. Therefore, the proposal is to implement the QDB as central search facility across all suitable CESSDA holdings. While the design of a single software application as 'the' QDB has not been the objective of WP9, the QDB is seen as part of a larger envisioned system where the resources held by CESSDA archives can be shared and reused. The distributed character of these archives demands a distributed architecture. An agreed metadata model is needed for integrating these. A design as used in relational database design will facilitate practical reuse and maintenance. The DDI3 metadata model uses this approach and will be a good basis (see also Alvheim, 2009).

A tender report has yielded a detailed model for implementing a technical infrastructure for this. In the following, the recommendations of the tendered report on QDB (Gregory et al., 2009) and its evaluation by WP9 (Hoogerwerf, 2009) are briefly summarized. The infrastructure should allow for:

- Locating questions through free text search and concepts;
- Linking questions to additional survey metadata / physical data / survey results;
- Linking from variables to questions;
- Querying for questions based on references.

On top of the goal of universal coverage, a QDB may also provide selective views on subsets of questions. These subsets may be extended – by e.g. manual entry through CESSDA staff or interested research groups – beyond the set of previously available questions, that is, they can be extended to include questions of special status or quality. This can, for example, concern sets of questions or questionnaires under discussion in international research groups, in the phase of questionnaire development, it could be questions that are recommended as 'gold standard' for certain measurement problems, there could be sets compiled for teaching and course work, etc. The entry, editing, grouping and sub-setting functionality require a separate web interface. It is easy to think of adding modular layers of functionality to that in a stepwise fashion as well. For example, an interface for organised question translation can be implemented at a later stage. This would build on previously included functions for grouping cross-nationally equivalent questions.

While the QDB needs a central search interface, its data holdings can remain entirely distributed across local repositories. In other words, according to the proposal of the tender report, it should make use of a registry, which keeps track of the contents of distributed repositories. This is structurally analogous to the current setup of the data catalogue on the CESSDA portal, does however use a more modern approach that will scale much better with increasing demands (Hoogerwerf, 2009). Through web services, resp. a service oriented architecture (SOA), local repositories shall actively register availability (and some metadata) of new question to the central registry, whereas the current CESSDA portal uses the local repositories in a more passive role, their contents being regularly harvested through the OAI-PMH protocol.

As a summary, regarding the design of QDB and its integration into the larger CESSDA RI, the tendered report on WP9 makes the following recommendations. These are endorsed by the evaluation report (Hoogerwerf, 2009):

- **Question Bank conceptual model:** Based on the DDI3 schemes, the QDB will act as a portal for available archived questions and as a question bank for newly created question (the latter to be achieved in the long term). As such, it will primarily function as a repository for locally held questions and as a proxy for access to non-local objects.
- **Architecture:** As outlined above, the general proposed architecture consists of repositories, a registry for efficient search with locating services and 3CDB facilities. All objects need to be uniquely identified using URNs.
- **Repositories:** As part of the architecture, the repositories represent the archives and act as single or multiple object banks. To ensure integrity and stability, the infrastructure should provide some redundancy, and CESSDA should implement quality-requirements (availability of online services) on participating data providers.
- **Registry:** the registry is meant to support building and maintaining relations between objects, such as between questions and variables, and the embedding of those within studies. At the same time, the registry provides a first layer of search functionality for the materials it indexes. How much of e.g. the searchable material needed for a QDB should be replicated in a registry – and thus be accessible for direct searches on the registry itself – has not yet been defined. At any rate, the registry must be able to forward QDB search requests to repositories with the full question text material.

- **Metadata specification:** the following metadata standards of particular relevance were named: DDI2/3, SDMX, and the ISO/IEC 1179 and ISO 19115. It has to be noted that DDI was designed to work together with all these standards. For example, combining SDMX and DDI could provide the ability to maintaining the linkage between survey data at the micro level and aggregated data.

One central observation must be made here: the proposed infrastructure is not only addressed towards a QDB, but is designed to become the backbone of all of CESSDA's future data services. This includes embedding all data related portal services and the 3CDB, as well as any other service that can be realised in a repository. Therefore, QDB (and 3CDB) are only possible test cases, and other cases such as including e.g. aggregate data services or services for privacy protected data can easily be thought of.

Open issues that have not been tackled by the WP9 team are those of communication protocols (beyond the proposals made in the QDB Tender Report) and of minimum content quality specification. It would be desirable that the QDB contains only question texts of a granted degree of completeness, also regarding the question context metadata, such as information on the flow of control, interviewer instructions, neighbouring questions etc. Such requirements will have to be detailed before the actual implementation of a QDB can be begun. Ideally, this will be resolved indirectly by sufficient CESSDA quality criteria for *any* piece of metadata made available through any CESSDA member's dissemination system. If CESSDA decides to differentially mark the quality status of its metadata holdings ('seal of approval' etc.), the ingest process of the QDB should be able to either reject submissions with insufficient quality status automatically, or to convey the markers to QDB end users unchanged.

If a manual quality checking process for questions is part of the final specification, this would require a QDB management team with substantively competent staff. If no such intellectual controls are required, the technical and organisational location of the QDB can be determined by technical criteria alone. It could, for example, be integrated directly into the CESSDA portal software. In contrast, there is no technical or organisational reason that the administration staff responsible for the entry, editing, and sub-setting functions reside physically close to the servers of the CESSDA portal. These tasks can be allocated to CESSDA members according to available resources and competencies.

However, the evaluation report also points to serious challenges implied in the architecture proposed by the tender report. These mostly relate to the sheer scale of the task of implementing a universal distributed infrastructure with high-availability services. In particular, two sub-tasks appear challenging, and their first implementations will need careful and extended testing: (1) a CESSDA metadata model that partly even overarches DDI 3.0 (by connections to SDMX and a query language), (2) developing a high-performance and high-availability registry that basically will have to register and index any single object – as defined per the metadata standard – that is held in the CESSDA online systems, and will have to manage references between all online systems.

Therefore, realistically implementation must be approached in a stepwise manner again. A first sketch can be found in the evaluation report (Hoogerwerf 2009).

3.2 Technical Requirements of QDB

Practically, the current CESSDA portal could already satisfy much of the search needs of the proposed QDB if two extensions were put into place: 1) Question texts were indexed selectively and made searchable at a level as implemented in stand-alone NESSTAR servers. 2) Searches and corresponding result sets could also be defined and retrieved in machine-to-machine interaction, and not only through a GUI. Therefore, the QDB can either be implemented as part of a complete web service-driven network as proposed by the QDB Tender Report, or as a probably straightforward extension of the current CESSDA portal. However, this extension of the current CESSDA catalogue would clearly not satisfy the needs of a DDI3 based infrastructure as described above. For this, the crucial points to fulfil are these:

- CESSDA will adopt DDI3, and archives comply with DDI3 metadata specifications;
- CESSDA is able to host and maintain the described large scale infrastructure;
- CESSDA is able to assume long-term responsibility for stability of metadata and data (persistent identification).

The second desired feature of the QDB is the ability to store additional question materials, beyond those already held in CESSDA as part of study metadata. This requires an editing interface and a user management system and is in this respect more demanding than the basic search functionality.

3.3 Best Practice Expectations for the QDB

As described above, the QDB ideally underlies the same mandatory requirements and best practice recommendations as other CESSDA metadata repositories with question text material. WP 9 does therefore not make very specific recommendations in this area. The items listed below can inform the general CESSDA discussion of metadata requirements:

- Metadata without at least literal question texts are useless for the QDB
- Literal question text plus response categories are the only *strictly* required components (plus study context and provenance via reference/PID)
- Desirable metadata:
    - Concept tagging of questions, using ELSST
    - Routing information etc. (addressed by DDI 3)
    - CAI actionability (addressed by DDI 3)
- Tagging question entries by completeness, possibly also by degree of quality control

3.4 Future Resource Needs of the QDB

Obviously, planning for the QDB is contingent on many decisions that have to be made outside WP 9, and it is intertwined with planning for the general data infrastructure. The evaluation report recommends a stepwise approach and estimates

an effort of between 15 to 30 person months to implement a proof of concept for the QDB in a web services infrastructure with a few pilot member archives. Deployment as a full scale infrastructure would require multiples of this in development and installation effort. It is certain that the maintenance demands of such an infrastructure will be significantly higher than those of the current portal and catalogue services.

4 Collaboration Options

The GE*DE project (a node of the DAMES network at University of Stirling) is building a data harmonisation system with some similarities to the 3CDB. In particular, the GE*DE system already has a working data processing engine (technically robust, but the interface is still at prototype level) for matching categorical data across sources. The GE*DE project group has indicated willingness to cooperate in community building and exchange of harmonised data materials, and possibly in joining software components. It should also be considered whether the GE*DE project group can become a partner in future applications for funding.

The proposed 3CDB approach is sufficiently general to harmonise data of almost any kind, not only individual level survey data. Therefore, the system could be immensely useful to NSIs and EUROSTAT in their efforts to harmonise official statistics across Europe. Further, the systems use could be extended to any aggregate data, which would make it useful to economists as well. Collaboration with NSIs should therefore be sought after a more mature proof-of-concept application has become available.

5 Policy Recommendations

Clearly, 3CDB and QDB are deeply intertwined with the overall technical infrastructure of the future CESSDA-ERIC. When they are developed to their full scale as proposed here, they could very probably become show case applications for the benefits that result from networking European resources not only organisationally, but also technically. At the same time, it is evident that the implementation of the infrastructure is a multi-year endeavour that requires careful planning and extended testing phases for interim steps. It is unlikely that this can be mastered in a single project, or even under a single funding programme and period. Rather, a roadmap for the implementation of the infrastructure should be developed as one of the first tasks for the new CESSDA-ERIC management.

At the practical level, follow-up projects for first steps towards the 3CDB and QDB should be implemented as soon as possible, yielding operational and publicly visible results of any new infrastructure option as soon as it becomes available.

However, some strategic decisions must be taken as soon as possible in order to credibly commit all CESSDA members to preparing for this long-term project. The first is that CESSDA should commit to a common metadata model based on DDI 3.0. This implies, secondly, that an infrastructure for persistent identifiers is established as quickly as possible, which probably requires action on part of all national members.

6 References

Alvheim, Atle (2009, August). *D5.3. A CESSDA Common Data portal.* Accessible via the CESSDA PPP intranet: *http://www.cessda.org/ppp/wp05/WP5_Common_Data_Portal_v.2.0.pdf*

Atkinson, T., Cantillon, B., Marlier, E., & Nolan, B. (2002). *Social indicators*. Oxford: Oxford Univ. Press.

Bourmpos, Michael; Linardis, Tolis (with Alexandru Agache, Martin Friedrichs, and Markus Quandt) (2009): D9.2 Functional and Technical Specifications of 3CDB.

CESSDA (2008-2009). *CESSDA PPP - Preparatory Phase Project for a Major Upgrade of the Council of European Social Science Data Archives (CESSDA) Research Infrastructure*. Online resource, last accessed 2009-05-18: http://www.cessda.org/project/

CESSDA PPP - Work Package 9 (2008). B*uilding an Infrastructure for Content Harmonisation and Conversion*. Online resource, last accessed 2009-05-18: http://www.cessda.org/project/doc/wp09_descr2.pdf

Gregory, A., Heus, P.; Nelson, C., and Ryssevik, J. (2009): *Technical Specifications for a European Question* Data *Bank. Tender Report to the CESSDA-PPP*, available at: http://www.cessda.org/project/doc/CESSDA_PPP_QDB_May09.pdf

Hoogerwerf, M. (2009): Evaluation of the WP9 QDB Tender Report.

King, G. (2007). An Introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods and Research*, 32, 2: 173-199.

Krejci, Jindrich; Orten, Hilde and Quandt, Markus (2008): Strategy for collecting conversion keys for the infrastructure for data harmonisation, http://www.cessda.org/ppp/wp09/wp09_T93report.pdf

Quandt, M., Agache, A., & Friedrichs, M. (2009, June). How to make the unpublishable public. The approach of the CESSDA survey data harmonisation platform. Paper presented at the *NCESS 5th International Conference on e-Social Science*, 24th – 26th June 2009, Cologne. Accessible at: **http://www.ncess.ac.uk/resources/content/papers/Quandt.pdf**

Related projects and platforms:

CCESD-IS: Centre for Comparative European Survey Data Information System http://www.ccesd.ac.uk/

GEODE: Grid Enabled Occupational Data Environment http://www.geode.stir.ac.uk/

RAMON - Eurostat's Metadata Server http://ec.europa.eu/eurostat/ramon/