Integrated metadata for the harmonization of a National Data Infrastructure

Susana Portillo Cruz[1]

[1]*Central Statistics Office, Ireland*

Abstract

National Statistical Institutes (NSIs) are facing increased pressures to disseminate timely, accurate statistics with users now demanding data and information faster, more frequently and at a more granular level. This requires National Statistical Institutes to invest in the establishment of a data ecosystem at national level to make full use of secondary and other new data sources which aren't always structured in a consistent, standarised way.

This paper discusses the efforts of the Central Statistics Office in Ireland to provide a technical solution for the use of harmonised concepts, questions and responses using international standards. This solution can be used by the NSI and government agencies collecting administrative data to ensure the consistent representation of variable values with the aim of using administrative data faster and more efficiently to achieve more timely and accurate dissemination of statistics.

Keywords: harmonisation, standardisation, National Data Infrastructure, Tools, Data Stewardship

1. Introduction

Data is one of the largest strategic assets for any modern Government since it facilitates evidence-based decision making on national policies and strategies. This approach to governing builds the foundations for public trust in those policies.

Modern societies are moving at a faster pace than ever, and our policy and decision makers currently need more frequent, timely and granular evidence and insight, at a time when National Statistical Institutes (NSIs) are experiencing a large fall in responses to traditional means of data collection.

Principle 1 of the European Statistics Code of Practice places NSIs in the role of coordinator of statistics within a country and plants the seed for this coordination at the level of the data collection and not just dissemination. By taking a leadership role in the creation of a National

Data Infrastructure (NDI) that will facilitate the use of administrative sources to access more timely and rich data NSIs will increase their capability to produce the required statistics at the required level of detail when they are needed.

The implementation of this infrastructure brings challenges such as the development, implementation, and governance of common data standards across the Civil and Public Service. These are the standards that should facilitate data linkage and integration across data regardless of the source however, operationally, different departments and even different areas with the same department make silo-ed decisions on the collection and storage of their data citing differences with everyone else.

This silo-based approach in turn stems from a lack of clarity, direction, and support where different departments do not communicate with one another leading to a lack of coordination into the system. In the implementation of principle 1 and as part of our role as Data Stewards within the Irish Statistical System the CSO has assumed the task of coordinating and disseminating data collection standards within public organizations in the Republic of Ireland.

2. Standards

Among the common standards to be implemented across the NDI are the usage of common concepts and common values for the collection of specific variables across the system, be it via administrative data or collected by means of a survey. We also have the issue of using common input and, in as much as possible output classifications across domains and departments. Implementing those standards generates greater interoperability of systems and data, improves the quality of the data across the system and maximizes the value of any data asset.

2.1. Current state & harmonisation

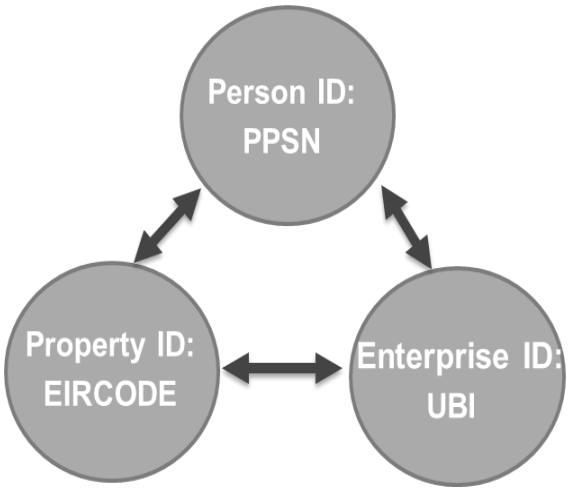
The NDI in Ireland, led by the Central Statistics Office (CSO) is in the process of moving from a silo-based environment where no standardization of data occurs to a system based on harmonized concepts and classifications. Our statisticians are finding it hard to meet the demands of policy and decision makers due to a lack of comparability between the concepts measured or collected by the various departments.

The first step taken towards standardization has already taken place in the form of using unique identifiers across the Irish Statistical System and other departments providing administrative data. Three types of identifiers exist within our national system:

- A unique person identifier used by all citizens to carry out official business termed PPSN (Public Personal Security Number)
- A unique business identifier used by enterprises and establishments for their operations termed UBI (Unique Business Identifier)
- A unique address identifier given to every single dwelling in Ireland, the EIRCODE

It is clear that there are common links between these three identifiers: an individual's PPSN can be mapped to their residence via the Eircode. Similarly, an enterprise operates from a premises that has its own Eircode and, finally, a person with a PPSN can be employed by an enterprise that operates under a unique UBI. What is not so clear cut is that, in some cases, the relationship is not a 1:1 one.

Figure 1: NDI Common identifiers

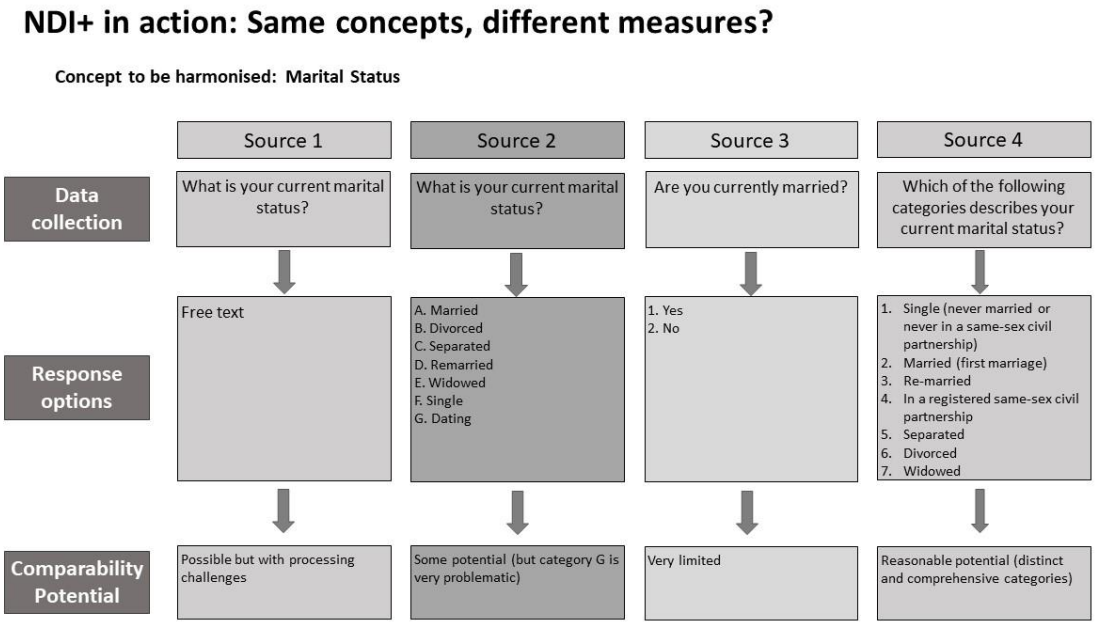


With the operation of unique identifiers in place, harmonization can be moved to sharing data standards such as standardized concepts, reference classifications and metadata as well as the use of harmonized questions, categories, code lists and outputs. This is the role that the CSO has integrated under the Data Stewardship concept and has been labelled as NDI+.

If we take the concept of ‘Marital Status’ for example we can see that currently there are multiple ways to, not only ask the relevant question of a respondent, but also to record a particular answer.

The lack of proper definitions for response options is yet another problem that affects comparability.

Figure 2: current state



While it is not possible to standardize every single element, we should still aim at having a level of harmonization where departments using technology to gather information can avail of, at least, information on a “preferred” of “standard” or “harmonized” manner to collect and disseminate the information without it being so prescriptive that it generates a multitude of special cases or dispensations.

Harmonization of classifications at the time of input then makes it easier to streamline classification in aggregated disseminated products and facilitates output comparisons thus ensuring users are comparing ‘like with like’.

2.2. Management and Governance

Determining these standards is not an easy task and requires an initial level of consultation with all departments collecting similar information to agree on the harmonized concept, classification and values. An NDI champions group has been established to govern and give direction towards harmonization.

Once agreed, the NSI in its role of coordinator should be the custodians and governors of these harmonized concepts and facilitate their access by the general public.

3. Solution

The current proposed solution to overcome the issues discussed so far involves two separate processes: that of managing the harmonization of concepts, code lists, questions in used instrument or variables in administrative sources and output classifications, and the process of storing this information in a system that allows for discovery and reuse of the determined standards. Each of these two processes is separately described under its own heading below.

An indispensable part of the harmonization process is to ensure the visibility, transparency, and accessibility of harmonized objects. Any solution devised will need to involve specialized software to store the information and linkage of objects in a dynamic database and the necessary tools for users to, not only be able to view these objects but also to be able to access them programmatically from such database regardless of any individual infrastructures at each department level.

3.1. Public interface

A bespoke public user interface has been created to allow for discoverability of harmonized standards. For each standard agreed, the interface will contain information on:

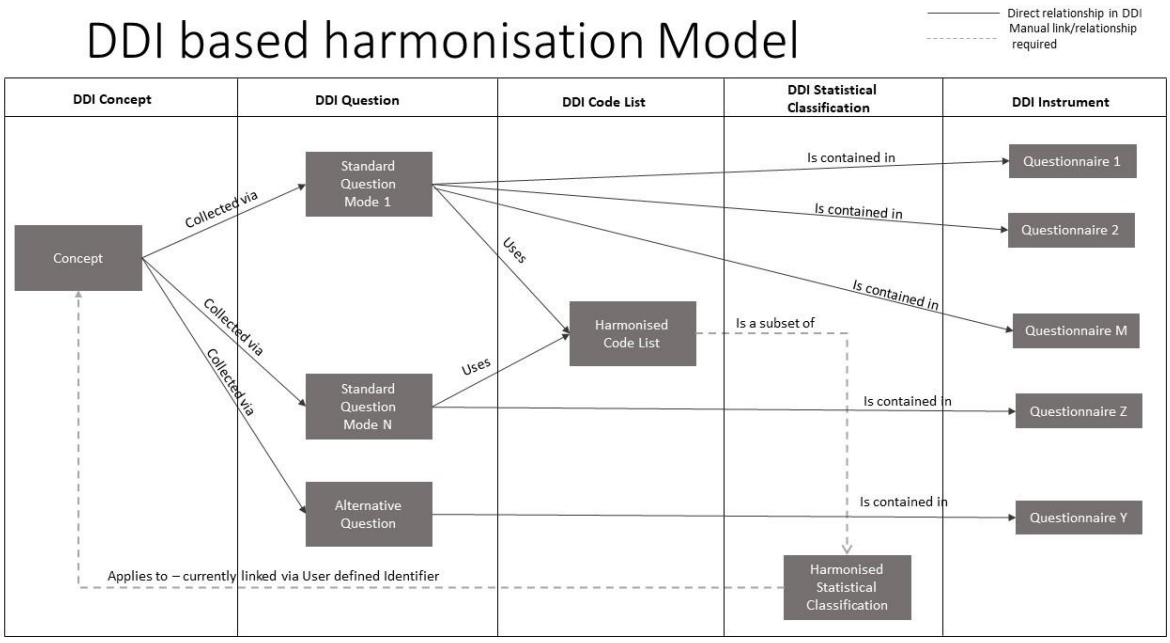
- The harmonized concept definition.
- The preferred manner to ask for information on the concept, including explanations on adequate collection mode.
- Alternative manners to collect information when the standard preferred manner does not suit users.

- The standard single-level code list to be used during collection.
- The multi-level statistical classification that can be used at dissemination time, explaining its correct breakdown usage and how to aggregate/disaggregate items as required.
- Examples of usage of those standards within the national statistical institute
- An API call that can be used by software developers to programmatically extract the latest version of any code list and/or classification and integrate them into collection systems, be it for building questionnaires or to record variable values in administrative data.

3.2. Model and storage

The most natural and feasible solution for CSO involves the use of the DDI standard via a Colectica repository given that the office has extensively used Colectica for questionnaire design and as a question bank for several years.

Figure 3: DDI based harmonization Model



The backend structure of the Data Standards Portal is based upon the 5 main DDI objects depicted in figure 2.

- The “Concept” object is at the center of this structure with all other objects revolving around it in a structured manner.
- Information on concepts can be collected via a “Question” object. There may be multiple standard questions to collect information on a concept depending on the mode of operation and any question deemed standard has a direct link to the concept it relates to.
- In some cases, where we know that exceptions can occur, we have allowed for the use of ‘Alternative’ questions. These questions still refer to the same concept however allow for a different way to input data than a standard question.
- Standard questions have a prescriptive way of collecting information as “Code List” objects. It is these code lists that present the most difficulty in harmonization, but it is where the most gains are to be had.
- A ‘code list’ in turn is a subset of a harmonized “Statistical Classification” object. The difference between them is that, in the standard, a classification allows for multiple levels of options or categories to group the information whereas a code list only allows for a flat representation of categories since it is more geared towards input data.
- Once these standard questions and code lists are in use via ‘Instrument’ objects, users will be made aware of where all these standards are used within CSO and will be able to make informed decisions on the usage of the standards.

The information and linkages are extracted directly from the Colectica repository using its proprietary software development kit and is built using .net technologies.

3.3. Adapting DDI

Because the project was based on a visual representation of the standard to make it user friendly, some customization of DDI objects within Colectica had to be implemented since the DDI standard did not cater for every single type of information that was needed. Because Colectica has been used extensively for several years there were already a wealth of objects in our repository. The need to differentiate Standard from non-standard objects led us to use DDI’s User Identifiers within the object as key-value pairs. Among these identifiers are:

- Type: “Harmonised Data Standards & Guidance”
- Rel_Type: “Preferred” for standard questions or “Alternative” for alternative questions.

- StatClassURN; contains the DDI URN of the statistical classification that serves the concept. This was introduced due to the need to link a statistical classification with a concept where there is no existing link within the Colectica System.
- ConceptGroupID: the Id given by Colectica to a Concept Group used to represent concepts that are related to one another.

User defined identifiers however cannot be viewed easily within the system when building Instruments without searching the discovery portal or the designer tool. To minimize disruption when using the standards and determining whether something is standard or not we also introduced a naming convention for objects that includes the abbreviation _STND in the name of each object created.

Similarly, Colectica creates a new version of an element every time a change is made, whether minor (e.g., a typo or errata) or major (e.g., a change in the categories and codes of a code list). To manage these types of revisions a “version” user identifier has been introduced within the object. The version number will be changed up only when a major change takes place within the object. Any major revisions will be governed by the Data Stewardship team and agreed with necessary stakeholders prior to

Markdown is used within the description of an object to separate different blocks of information to be displayed in the Data Standards Portal via a CSS. In some cases, extra information is needed. Extensive use is made of items such as “Instructions” within questions or even “Annotations” for certain objects.

3.4. Governance of the technical Solution

When a new standard is created (or an existing standard is updated) by the NDI champions Group, a dedicated team of Colectica administrators create the objects, structure and linkage among those objects within the Colectica repository. The software behind the portal is then smart enough to pick up any new or changed information and display it in a standardized manner for all users to access.

The standards are built on a development repository until such a time that all required elements are present and synchronized with one another. It is then that the standard elements will undergo an approval process managed by the metadata team to be released to the general public and a new version accessible via the portal and APIs.

4. Difficulties

The implementation of the proposed solution comes with inherent difficulties, some of those already encountered and others that are envisaged to arise during rollout.

Among the already managed difficulties are the technical aspects of developing the application. The public facing site requires programming to read metadata stored using a DDI standard and stored in an existing database with a proprietary model. Although the vendors of the original Colectica repository provide a Software Development Kit as one of their products, the learning curve for someone with no prior knowledge of DDI or Colectica is a steep one to be able to programmatically identify the elements required for easy representation in the portal. This element had the impact of slowing down the timeline of the development process. It was however offset y the even slower process of agreeing on the standards as both ended up finalizing on a similar timeline.

A second type of difficulty we envisage but one that has not yet occurred due to the standards not being fully established is a behavioral one. The integration of standards into processes and software requires not only a change of mindset to move away from siloed approaches but also modification of existing systems to adapt to said standards. This will inevitably require a change in the processing environments by modifying existing programs, be it the statistical programs within the NSI or the software used in departments to collect the administrative data. And this is no easy feat with years of legacy programming needing to be changed. The current silo mindset of individuals needs to take a shift into recognizing the overall benefits of the harmonization approach for everyone involved for the envisaged harmonization approach to be successful.

5. Conclusion

Using common standards at a national level can assist our statisticians unlock the potential of the country’s data assets. This harmonization should increase coherence and comparability of our data and the ability to combine data from many different sources so that more timely and granular statistics can be provided to facilitate policy and decision making. The standardization process is a very slow journey with multiple stakeholders involved along the way. It requires active engagement and ongoing consultation with all involved.

The software solution provided is a step towards encouraging all departments to adopt and implement these standards over time. It helps make the information transparent and accessible to the general public and supports the long-term vision of data linkage implementation at national level so our public bodies can fulfil the demands of our users.

Acknowledgements

We wish to thank our colleagues in the IT department in CSO for their invaluable input into creating this application in a relatively short period of time, given their lack of knowledge on both the DDI standard and the Colectica SDK.

On that note, we extend our thanks to the Algenta Colectica technologies team who have provided timely support to our IT team throughout the lifespan of this project.

References

DDI Lifecycle 3.3, https://ddialliance.org/Specification/DDI-Lifecycle/3.3/

Colectica SDK, https://docs.colectica.com/sdk7/

Public Service Data Strategy, https://data.gov.ie/uploads/page_images/2019-01-03-110200.740673Public-Service-Data-Strategy-2019-2023.pdf

National Data Infrastructure Ireland, https://www.ops.gov.ie/actions/innovating-for-our-future/data/