LUNDS
UNIVERSITET

AN ONTOLOGICAL APPROACH TO CONSIDERING METADATA ACROSS SPATIAL AND NON-SPATIAL DATASETS

Alannah Finn

Department of Human Geography
SGEM08
Spring 2021

Examiner: Nicklas Guldåker
Supervisor: Ola Hall

Abstract

This study presents the case for taking an ontological approach to metadata analysis as a means of facilitating enhanced interoperability across spatial and non-spatial datasets in GIS. A detailed literature review was undertaken in order to understand metadata definitions, concepts and usage, particularly as a "primary interoperability enabler" (Danko, 2008). Interoperability, within the context of this study, is understood as the ability to combine several types of datasets, specifically, demographic, health and spatial. The concept of ontology, which deals with the nature of being, is both presented and validated within the Theoretical Framework as a qualitative or descriptive approach to metadata analysis.

A detailed case study of the Demographic and Health Surveys (DHS) Program portal was undertaken, covering the full scope of the DHS Survey Process, with the purpose of identifying two suitable candidate countries for comparison. A further evaluation of the actual metadata categories associated with both sets of datasets was then performed, in order to navigate the sub-levels of metadata layers, facilitating the closer investigation of interoperability capacities within and across both candidate countries.

It was found that a flexible approach to an ontological framework involving three levels of application was effective. Furthermore, interoperability was found to operate within a single dataset, between specific related datasets and across the full range of dataset types. Through the consideration of appropriate metadata categories, both the context and meaning of such interoperability was better understood. Finally, applying an ontological approach, illustrates how the DHS's specific taxonomies effectively encapsulate deeper contextual meaning which cannot normally be expressed semantically in metadata tagging and layering.

Keywords: Metadata, ontology, interoperability, spatial, dataset, GIS, DHS Surveys

Word Count: 17,007.

Acknowledgements

I would like to express my upmost gratitude to my supervisor, Dr Ola Hall for his constant advice, encouragement, and support throughout this process, beginning in his GIS class, in his role as teacher and mentor. I would also like to thank Dr Ilkin Mehrabov at the Department of Strategic Communication in Lund University for his initial encouragement and advice on pursuing a topic within the area of Geoprivacy and locational technology.

To Prof Nadine Schuurman, thank you for your extensive and pioneering contributions to the fields of Geography, Critical GIS, and Metadata. Both your academic work and personal story are truly inspirational, which have helped me reach this point.

To my friends, teammates, peers, and roommates (Anne, Nikol and Sally the dog), thank you for standing by me in what's been an unpredictable and challenging past year. Your support, encouragement and kindness have directly contributed to the making of this thesis.

Finally, I dedicate this thesis to Dad, Mam and Oisín. Níl aon tinteán mar do thinteán féin.

Contents

List of Figures

Chapter 1 – Introduction

1.1 Aim, context, and motivation

The aim of this study is to investigate and understand the link between metadata and interoperability. Broadly, within spatial practices, metadata has a low utilisation rate, usually accounted to poor formulation and limited scope of its uses held by users, e.g., within GIS, metadata is generally viewed as a data descriptor function to be used in a catalogue program (Schuurman, 2009b). Additionally, spatial users, namely geographers, tend to avoid analysing metadata in a qualitative manner, or, in many cases, avoid interaction with metadata completely (Schuurman and Leszczynski, 2006, 716). Thus, for this study, the potential lies in investigating metadata application and usage in a mixed methods approach, i.e., both quantitatively (or semantically) and qualitatively (or ontologically). Therefore, the first principle aim of this study is to examine the importance of metadata in research, namely, Geography, importantly, asking how metadata can help in facilitating geographers utilise spatial and indeed non-spatial datasets. To further this, the second primary aim of this study is to investigate the role of metadata in facilitating interoperability. Interoperability in this instance, meaning the capability of combining spatial (e.g., GIS datasets – shape files, point data, etc) with non-spatial datasets (e.g., statistical spreadsheets, questionnaires, etc).

It should be noted that the goal of this study initially started out very broad, with the intention of investigating metadata within its whole lifecycle, i.e., from formulation through to standardisation and publication practices, in a generic (i.e., not case specific, such as DHS metadata categorisation procedures). Furthermore, the primary objective of the study would then have been to consider proposing a generic, process model which could act as a prototype framework for analysing metadata categories and fields across platforms. This scope was narrowed significantly and through examining the DHS case study, and now the focus of the study is to investigate the degree of interoperability that can be utilised, using the DHS's own proprietary meta process as the guiding framework. Furthermore, the breakthrough for this investigation comes through applying a multi-knowledge system approach to examining metadata categorisation, structures, and implementation.

Thus, the objectives of this study are as follows: firstly, to find a suitable theoretical framework which explain the concepts being investigated, highlight the gaps

within the literature and most importantly, provide a foundational framework for implementing the methodology. Secondly, to find a real life, working dataset portal to examine, preferably with a research background (i.e., peer reviewed), and notably, access its datasets and operationalise the theory. Thirdly, this study aims to investigate the nature of metadata, and thus its role in aiding geographers in navigating and utilising both spatial, and most notably, non-spatial datasets. Finally, this study aims to generate findings which can make recommendations, offer critique, and further contribute to the field.

1.2 Research question

With this addressed, the research now asks the questions:

1) How can an ontological approach to metadata facilitate interoperability across spatial and non-spatial datasets?

and

2) What benefits could geographers gain by employing an ontological approach alongside established semantic approaches to metadata?

1.3 Thesis Outline

The thesis is structured as follows: Chapter Two conducts a literature review of the key concepts, discussions, and studies within the fields of Critical GIS, ontology, metadata, geoprivacy and case studies specific to the utilisation of spatial datasets and using datasets provided by DHS. Furthermore, it outlines the main theoretical framework implemented in this study: an ontological approach to metadata, supported by concepts within Critical GIS. Chapter Three outlines the study design and discusses the DHS Program. In this chapter, the complete process of applying to and accessing DHS datasets is delineated and critiqued, while also discussing the DHS's own methodological practices, namely, the DHS Survey Process. Further commentary on the dataset's operational processes and metadata makeup is also provided. Chapter Four presents the key outcomes of conducting the study, examining specifically how an ontological approach to metadata facilitates interoperability within and

between datasets, while also critically analysing the DHS's metadata categories, data variables and specialised taxonomies. Chapter Five of this thesis presents and discusses the key findings of the results, further analysing the role of metadata categories, the degrees of interoperability and using an ontology-based approach to conduct the study. It additionally looks to further work and limitations. Chapter Six presents the key takeaways and conclusions from the research conducted and how they may relate to the broader field.

Chapter 2 – Theoretical Framework and Literature Review

2.1 Critical GIS: Ontological approaches

GIS has continually grappled with its place within the discipline of Geography, with traditionalists opting for approaches to remain dogmatic in conjectural framework and debate, avoiding any engagement with emerging technologies or adaptation to GI science applications (Longley et al., 2015, Schuurman, 2000). Dubbed as the "protagonists" within the field, critical GIS scholars aim to address the balance within this debate (Longley et al., 2015, 27). The field of Critical GIS seeks to present, enable, and critique the merging of social science theories with geographic information science (Schuurman, 2009a). In its more recent waves, it also focuses on the democratisation of such technologies, in addition to evaluating its interoperability, i.e., develop and espouse complex semantic structures with abstract theoretical concepts (Schuurman, 2009b; Schuurman and Leszczynski, 2006). Thatcher et al. (2018, 4) argue that GIS is a field that should never be so narrowly tied solely to a desktop. O'Sullivan (2006, 783) candidly encapsulates this sentiment, "If I were advising a new graduate student on how to succeed in geography these days, my advice would be to try to straddle the fence", the fence the author argues in this case being the fine ideological line between GIS and more broadly, the field of Human Geography. Schuurman (2009a) further notes that this epistemological divide must in turn be delicately balanced in capturing the disciplinary framework, i.e., clearing and critically defining the 'blend' of social processes and GI science.

It should be noted that due consideration has been given in the writing of this thesis, both to the positionality of the researcher and that of the dataset's gatekeeper, the DHS (see 3.2/3.3.2), in addition to the study's epistemological formulation, i.e., the life cycle of the DHS's datasets and the knowledge process (see 3.3.2). This study employs an ontology-based framework of metadata categorisation; thus, this section examines the fundamental concepts of ontology, through the lens of Critical GIS.

When considering ontological perspectives from a Critical GIS perspective, it is important to note the epistemological discourse which has emerged in the development of the

field. Critical GIS scholars have long operated from the perspective of an anti-positivist, pro-qualitative and bottom-up approach (Elwood et al., 2011; Pavlovskaya 2006; Schuurman, 2000). Pavlovskaya (2018) argues that the shift in positionality of Geographer's attitudes toward GI science and its systems has enabled it to be a tool for "social transformation". Expanding on this, the development of methodological approaches, i.e., ontologies, to adapt and even favour unorthodox approaches such as qualitative, participatory, and ethical GIS, further supports the contention by Pavloskaya (2018, 41, 44) that GIS technologies can move toward more progressive mapping consensus and further prioritise social change within research. Schuurman (2000, 571) adds to this, arguing that the previous "epistemological privilege enjoyed" by many science fields, including Geography, have been pushed to cooperate with their foes, and ultimately engage in a knowledge exchange. Thus, utilising geospatial technologies armed with the knowledge of social processes is the precise ontological purpose of Critical GIS approaches, "… GIS is designed to be used in conjunction with knowledge rather than a substitute for it" (Schuurman, 2000, 572).

In retrospect, the nature of metadata is entirely complimentary with that of ontology. Metadata serves to bridge the gap between the abstract knowledge processes employed by researchers and the complexities of data semantics, by constructing taxonomies (Schuurman 2009b; Schuurman and Leszczynski, 2006). Smith and Mark (1998, 308) define ontology as "the nature of being". The perspective of ontology through a broader geospatial lens, looks to build data exchange standards, which better develop human-computer interfaces for geographical information systems (Smith and Mark, 1998, 308). Using an analogy of a geological map, Schuurman (2009b) notes that ontology, as a concept, has been adapted for geospatial science as "… the way of seeing the world" (i.e., the map), through a classification system (i.e., the map's legend). Further, ontology serves as the gateway to interoperability across languages, community, and cultures (Goodchild, 1992, 10). Building on this, Schuurman (2009b; 2005) observes that epistemological consideration should be made in regard to the nature of metadata and its ontological purpose. There can be epistemological clashes among researchers when defining the ontologies captured within a field, e.g., the purpose of a landscape differs significantly between a logging company and an environmental activist, based on their worldview, notes Schuurman (2009b). On the complexities of categorisation, Schuurman (2005) observes that the definition of 'objects' is a notable example of the discourse among geospatial metadata users. Depending on the ontological perspective of the user, an object, in a computational sense, can be defined within strict

semantic classes (e.g., cartographic features such as point data), while others, applying a more abstract view, may characterise said point data as a house or a tree (Schuurman, 2005, 19, 21). Consequently, both viewpoints are correct, but require preceding consideration as to their contextualisation, an aspect which is discussed further in 2.2 (Elwood et al., 2011). Thus, the ontological capturing of metadata will always differ between the researcher and the user, depending on their epistemological perspective, and therefore should be acknowledged by scholars, of whom should act as "reflective observers", when applying ontological methodologies (Schuurman, 2009b; 2005, 21; Elwood et al., 2011).

2.2 The nature of Metadata

Metadata is, in its most simple definition, data about data (Schuurman, 2009b; Schuurman and Leszczynski, 2006; Danko, 2008). The Encyclopaedia of GIS defines geospatial metadata as data about "spatial information concerning objects or phenomena that are directly, or indirectly, associated with a location relative to the Earth; auxiliary information that provides a better understanding and utilization of spatial information" (Danko, 2008). Further, metadata is a "primary interoperability enabler", interoperability meaning the ability to merge several types of datasets (e.g., demographic, health and spatial) and utilising the same file or program formats (Danko, 2008). Further, metadata can take any form, i.e., electronic, or paper-based, while the 'data about data' can take multiple forms also, e.g., semantic data, numeric values, descriptive labels, etc (Danko, 2008). Typically within GIS use, metadata is required in at least four different cases: within geospatial datasets (e.g., shape files), historical records (of which provide legal standings to the use of geospatial datasets), in a user-decipherable form (i.e., processable and readable by humans as well as computers) and catalogs (Danko, 2008). Catalogs are of particular importance in GIS operation. Metadata is typically stored in the form of a statistical text file, which provides users with information about the data through search formats on location, author/producer, date, resolution, scale, and data structures (Danko, 2008; Schuurman and Leszczynski, 2006, 723). Most notably, metadata catalogs separate data-description from actual spatial analysis and processing, e.g., through subprograms of GIS software such as ArcCatalog (Schuurman and Leszczynski, 2006).

As noted previously, the definitive goal of metadata in a GIS context is to facilitate interoperability of data platforms. This is enabled by core aspects such as compatible

technologies or formats, as well as a common architecture (i.e., the data's file format can be joined or read by multiple platforms) (Danko, 2008; Schuurman and Leszczynski, 2006). This is only achievable through standardisation of the technology. Initial standardisation processes began as a localised process, with organisations such as the US Federal Geographic Data Committee (FGDC) developing the Content Standard for Digital Geospatial Metadata (CSDGM), enabling spatial data producers to universally encode spatial datasets about their descriptive nature, thus creating a formalised standard (Schuurman and Leszczynski, 2006; Smith and Mark, 1998). This process is now applied internationally, overseen by the International Organisation for Standardization Technical Committee for Geographic Information Standards (ISO/TC $_{211}$) (Danko, 2008; Schuurman and Leszczynski, 2006; Schuurman, 2009b). The ISO TC $_{211}$ uses metadata approaches as the mechanism for allowing geospatial data produced by one user to be semantically read and operationalised by another on an integrated platform (Danko, 2008). Many of these practices have been adapted from computer science approaches to standardisation of metadata. Using the same principles as mock-up languages for computer and web-based platforms (e.g., WWW, HTTP, etc), the meta categories for geographic data must incorporate a standardised documentation of the data's characteristics, e.g., individual geographical features, geometric attributes, etc (Percivall, 2008; Danko, 2008; Schuurman and Leszczynski, 2006; Smith and Mark, 1998).

This standardisation must be able to keep up with the evolving enhancements within GIS and other spatial technologies. Such standards must also address the commercialisation of GI techniques. The Open Geospatial Consortium (OGC) has further developed framework for interoperability across geospatial programs, most notably on web-based platforms (Percivall, 2008). The framework outlines regulation for key areas of commercial GIS use, categorised through tiers: the client tier, the application service tier, the processing services tier, and the information management services tier (Percivall, 2008). Individuals, organisations, and business's intellectual property is therefore addressed and importantly, protected under the framework, further enabling metadata to be openly shared and thus, achieving interoperability for users (Danko, 2008; Percivall, 2008). This element is important to note, as the DHS, this study's chosen case study, is an internet-based data portal which utilises commercial GIS data and map packages for its frontend clients, thus, standards for interoperability are a key component in its functionality (see section 3.2).

On the surface, employing a standardised modelling language, or following regulations for application methods to produce a prototype defining abstract theoretical

concepts, not limited to behaviours, relationships, and objects, (see 2.1 discussion) is the epitome of metadata (Danko, 2008). In layman's speak, metadata attempts to categorise the process of capturing real world knowledge and phenomena into a technical information system (Danko, 2008). Standardisation, while essential, is only one key component in the overall operability of geographic metadata. Schuurman and Leszczynski (2006, 714-715) note that the standardisation of metadata, is inherently hierarchical, meaning that standardisation of one area, e.g., classification methods, does not conveniently filter down to solve issues within other subfields, such as data integration, particularly at a semantic level. Geo databases rarely include detailed denotation on subfields like attributes, or indeed their intended use (Schuurman and Leszczynski, 2006). Previous metadata frameworks that have emerged from standardisation initiatives only allow for a "superficial" level of interoperability as they rely on the sharing of preestablished, institutional knowledge (Schuurman and Leszczynski, 2006, 715). Furthermore, many current metadata frameworks exclusively emphasise the geometric attributes (i.e., spatial, and numeric values) of spatial datasets (Schuurman and Leszczynski, 2006). This intrinsically narrow scope of metadata, derived from "the limited concept of the purpose of data documentation", often remits GIS practitioners from greater consideration for metadata's versatility (Schuurman and Leszczynski, 2006, 715).

Reiterating this argument, critical GIS scholars from feminist perspectives argue that the narrow scope availed to data semantics, i.e., categorisation and labels, only exacerbates the complex external political and social inequalities (i.e., the status quo) purportedly being examined through studies within standard GIS approaches (Elwood et al., 2011; Pavloyskaya, 2009; Kwan, 2007). Additionally, other critical GIS scholars highlight GIS's limited scope in contextualising its data (see 2.1). Schuurman and Leszczynski (2006) aim to address this, in their study on building an ontology-based approach to metadata. As previously outlined in 2.2, ontological framework is considered a foundational principle in Critical GIS approaches. Schuurman (2005, 19) comments, "Ontology … deals with the nature of being" of which the GIS scholars have used "… to address the problems posed by categories" (i.e., metadata). The key issue with attempting to apply ontological assumptions to metadata is the seemingly challenging aspect of combining semantic attributes with context (Schuurman and Leszczynski, 2006, 711). That is to say, the interoperability, or metadata functionality, has always operated via automation of the semantic integration through standardisation of semantic terms (Schuurman and Leszczynski, 2006, 711). In other words, the 'data dictionary definition' as dubbed by the authors (2006, 711), of semantic terms and field names are set in

stone, with no suppleness as to their connotation, while the reality within the field, or "on the ground", is adapting such terms with new interpretations and meanings. It is precisely this phenomenon which brings the authors to introduce a framework developed for constructing new metadata categories using ontological information:

Field	Description
Sampling methodologies	indication of *how* data was collected
Definition of variable terms	naming conventions, etc. used to identify and describe entities and attributes
Measurement specification	measurement systems; instrumentation; thresholds as well as range (e.g. clarification of the maximum and minimum)
Classification system	documentation of classification scheme used and taxonomic details; this is the *collection* of variable terms
Data model	specification of the data model, including history – for example, have different data models been used in the past?
Collection rationale	logics behind data collection and the digital encoding process – an indication of *why* data were collected in the way that they were
Policy constraints	legal or other constraints or influences on data collection or classification
Anecdotes	additional comments pertinent to understanding how to use the database; not to be confused with *abstract,* an existing field

Figure 1: Metadata fields for capturing ontological information

Source: Schuurman and Leszczynski (2006, 718).

Using ontological conceptualisation, the framework allows geographers to create their own contextualised meta categories, formulated through broader epistemology, e.g., field observations, methodological approaches (e.g., survey collection – see 3.2), etc, as outlined in the table above (Schuurman and Leszczynski, 2006, 718). This framework further aids researchers to carry out the careful evaluation of potential datasets, in a deeply contextualised manner (Schuurman and Leszczynski, 2006, 716). This aspect is an essential element regarding the application of this study, as is further discussed in chapter 3 (see 3.1-3.2). Furthermore, the framework is designed to assist researchers in creating a metadata taxonomy for potential data collected or indeed generated, in a way which can optimise its usability, and thus, the interoperability, by capturing the "deep context" of datasets through the database description, or the catalog function (Schuurman and Leszczynski, 2006, 716-

717). This is retrospective of the reality that most spatial data users do not engage with potential gradations available through a descriptive process of analysing metadata (Schuurman and Leszczynski, 2006, 716).

The authors stress that while the framework is adoptable, the nature of the framework's applicability and thus its interoperability, is seemingly case-specific (Schuurman and Leszczynski, 2006, 723-724). The implementation of ontological approaches to capture meta fields, operationalised through epistemology, truly reveals the complex nature of incorporating the reality of the chaotic spheres of the world with the outwardly rigid semantic integration (Schuurman and Leszczynski, 2006, 723). Be this as it may, the ontological framework outlined aims to arm geographers with a guiding structure using the conceptual tools from philosophies (ontology and epistemology) of which they are best qualified (Schuurman and Leszczynski, 2006, 723). It is indicative of this, that this framework acts as the key theoretical foundation in conducting this study, as is applied appropriately, and discussed further in Chapter Four.

2.3 Geoprivacy and Geo-Ethics

It could be argued that the emergence of geo-ethics has been a direct consequence of Critical GIS and its subdisciplines. While ethical concerns within the practice of spatial technologies have long existed as an established discourse, the presence of geoprivacy as a more 'mainstream' subject matter in circles outside of Geography, has coincided with the evolution of many academic disciplines in the 'big data' era (Swanlund and Schuurman, 2019; Longley et al., 2015; Diamond et al., 2009). Furthermore, societal opinions, politics and attitudes have also seen a significant shift in mentality, with recent global events such as the ongoing Covid-19 pandemic giving rise to questions about the use of spatial and location-based technologies, as a means of protecting society or policing it (Calvo et al., 2020). While the implementation of ethical frameworks for GIS and geospatial technologies are outside the scope of this study, geoprivacy as a continuously emerging field, has had an active role during the course of this study. Further, from a critical perspective, an ontological review of existing debates and practices within such fields can aid in epistemological development of GIS and geospatial practices. Most notably, the assessment of such literature has also aided in providing this study with a comprehensive understanding in spatial and locational data

privacy and ethical issues, an important element when selecting the candidate portal (see Chapter 3, section 3.1 and 3.2). Thus, this section offers a brief synoptical review of upcoming areas within the sphere of Geo-ethics and Geoprivacy that are of interest to this thesis.

Geo-ethics has been an area less considered compared to many cross-disciplinary subjects within Geography, with few notable exceptions, such as Feminist Geographies (Pavloskaya, 2018, 43). Much of the emerging framework for ethics and 'good practice' principles have come from work within new emerging areas such as Participatory GIS (PGIS) (Rambaldi et al., 2006). As PGIS seeks to work on the ground with communities through volunteer work and community engagement, this has necessitated the need for framework and guidelines to emerge (Rambaldi et al., 2006). Further, the emergence of knowledge-sharing platforms such as internet forums for practicing PGIS researchers, has developed both institutionalised and voluntary-lead groups such as the Research Data Alliance (RDA), PPGIS Network, the Open Forum on Participatory GIS, and the Geospatial Professional Ethics Project, among many. This in turn, has seen ethical-guidance frameworks develop as a result of their work, e.g., *Participatory GIS (PGIS) practice* (see Rambaldi et al., 2006). Furthermore, as such practices move from the academic to the commercial sphere, there is a growing call for geo-ethics to be wider taught at post-graduate and notably undergraduate levels. Sinton (2017) notes that some institutions may devote as little as a single lecture to the topic of 'ethics' in their introductory GIS classes, arguing for changes to syllabi, which ought to reflect the vast number of subtopics (e.g., mobile data tracking, geomasking techniques, etc) and their practices, taught "… regularly, not just once and in Week 12".

The realm of Geoprivacy has developed from the prevalence of Location Based Services utilised by governments and private digital patents, in addition to the emergence of mobile data tracking and the sharing of spatial information on digital platforms (Weiser and Scheider, 2014; Nouwt, 2008). Furthermore, the concept of the 'Geoweb' (geospatial web) examines the new kinds of geographic information available, coinciding with the development of Web 2.0 (i.e., social media platforms, embedded applications, etc) such as Google Street View, Snap[chat] Maps and TwitterGeoAPI (Elwood and Leszczynski, 2010). Weiser and Scheider (2014) present three key stages in the development of the Geoweb phenomenon: the creation of the digital persona (first developed by harvesting personal information for geomarketing), tracking (the gathering of GPS location data [or 'movement data'] and further assessment of activities at precise locations) and "knowing (almost)

everything" (the combing of personalised location history with other forms of personal data - an example of interoperability). The privacy entitlement and data protection laws governing such phenomena is, to this day, at the front of scholarly discourse. While recent developments in data privacy laws, such as the European GDPR, have bolstered data privacy advocation, the difficulty lies in keeping pace with technological developments, most notably, when it is private patents such as Google and Facebook, who 'create the rules of the game' (Ferretti et al., 2020; Elwood and Leszczynski, 2014; Nouwt, 2008). Thus, governing bodies and regulators are continually one step behind the constant evolving of data harvester's techniques to secure and potentially exploit geo-information (Nouwt, 2008).

Geosurveillance is a significant field within Geoprivacy. Evolving from GIS's military roots, geosurveillance mechanisms uses location tracking techniques as a means of data collection. Scholarly consensus varies as to the benefit or indeed the consequences of such practices (Swanlund and Schuurman, 2019, 598). A notable example is the Smart City movement, a familiar topic across Geography and other disciplines. Geosurveillence critiques argue that from a data perspective, tracking methods used in smart city application has resulted in exhaustive surveillance on movement and even behavioural data (Swanlund and Schuurman, 2019). In addition to critique of approaches, scholars also argue that resistance to geosurveillance is becoming increasingly more difficult, particularly for private citizens (Swanlund and Schuurman, 2019). Many techniques, such as International Mobile Subscriber Identity (IMSI technology) catchers are particularly intrusive, as they rely on the ostensibly 'trivial' fact that private citizens are heavily reliant on their mobile (cell) phones (Swanlund and Schuurman, 2019, 596-597). Further, such techniques are difficult to avoid, as they are employed by private patents (e.g., cell phone operators) and even governments (see Swanlund and Schuurman, 2019). Such practices have been implemented throughout the ongoing Covid-19 pandemic, notably at the beginning of large European outbreaks in mid-2020. Mirroring tracking techniques deployed by their Asian counterparts in the early stages of the disease outbreak, European researchers looked at private citizen's cell phone operator records, as a means of tracking and analysis the general public's location movement during the imposed 'lockdown' periods throughout Europe (Ferretti et al., 2020). Additionally, individual location data records were also analysed for the same geographic locations, provided by social media sites such as Facebook (Ferretti et al., 2020). In Israel, researchers found health authorities to take this one step further, replicating geosurrveillance techniques previously used by counter-terrorism units, as a means of effectively tracking mass and local in outbreaks of corona

(Calvo et al., 2020). Such technical deployment has given rise to an emerging field in 'digital epidemiology'. Regardless of the epistemological beliefs of geosurveillance critics or practitioners, Kostkova (2018) argues that the interoperability between technologies and their uses has triggered a new research "frontier" locational data collection and analysis. On the ever-invasive expansion of geosurveillance methods, Dobson and Fisher (2003) go so far to argue that we, as a society, have enter a new paradigm, provocatively coined 'Geoslavery'. The authors (2003, 47-48) define this as being the practice "… in which one entity, the master, coercively or surreptitiously monitors and exerts control over the physical location of another individual, the slave". In essence, society's new form of slavery, is that by location control (Dobson and Fisher, 2003, 48).

2.4 Case Studies

This section examines case studies which use the Standard DHS survey datasets as the main data source for the study. While the key focus of this thesis is to analyse the metadata categorisation of the DHS dataset files, rather than utilise the raw datasets, this section provides a comprehensive illustration of how the DHS Standard survey, notably the spatial and multifaceted datasets, are employed in the real world.

Section 2.5.1 examines core literature that analyses techniques in geomasking on the geographic coordinates within Standard DHS Survey datasets, of which are interoperable with highly sensitive data such as health and commercial datasets. The chosen literature discussed illustrates the sensitivity of the DHS survey datasets in relation to participation's geographic location and the care that ought to be afforded when using them, by presenting reversibility techniques. Most importantly to this investigation, the authors further highlight key issues regarding the use of spatial data by non-spatial users.

Section 2.5.2 assesses a publication, distributed through the DHS, on use of their Spatial Covariates datasets as a means of measuring health issues against environmental factors. The Spatial Covariates, as discussed further in section 4.2, embody full metadata functionality, or interoperability, by merging spatial data such as geographical coordinates, with non-spatial datasets, i.e., the demographic and health (biomarker) datasets.

2.4.1 Reverse Engineering

While the area of Geoprivacy and Geo-ethics has expanded significantly in the past decade with the emergence of location-based services through digital platforms, a significant reflection within this study is the use of GIS: geographical techniques, the datasets used, the techniques involved to 'mask' participant's personal information (geomasking) and the potential implications for both researchers and the participants. The Demographic and Health Surveys (DHS) datasets have been at the forefront of many academic debates in relation to the issue of displacement (Seidl et al, 2018; 2015, Grace et al., 2019, Dorélien et al., 2013). The datasets are a collection of individual and household characteristics – ranging from health data to population index, collected throughout countries in the third world (Dorélien et al., 2013, 415). The premise of collection of such data in a spatial formation (i.e., GIS point data or coordinates), allows researchers to build a spatial survey cluster, thus illustrating the condition of certain villages or towns based on health, environmental condition, wealth index, etc. DHS datasets use geographic coordinates (geocodes) for each cluster, which make up 15-30 households, characterised as rural or urban settlements (Dorélien et al., 2013, 415). This data can then be linked with other datasets, such as aerial images, or health indicators, depending on the research objective (Grace et al., 2019, Dorélien et al., 2013). The issue arises within the potential identification of survey participants through locational or other demographic factors provided in the survey responses. To protect the identity of the spatial clusters (villages) and therefore the participants, the coordinates within the datasets are shifted and displaced by 0-5km from their original collection point, with a further 1% of data randomly displaced by up to 10km (Grace et al., 2019).

Most notably, evidencing an understanding of core geographical concepts is essential, as identified by researchers, when working with datasets such as these (Smith and Mark; 1998, 316). Grace et al (2019) use spatial population data from the DHS, combined with climate data taken from another source, to investigate the environmental profile of the settlements of their chosen study area using a case study analysis (two different mountainous regions in Tajikistan, sharing almost identical spatial, environment and demographic characteristics). While replicating the core methodological processes (geomasking techniques) are widely outside the scope of the methodological considerations of this study, the key considerations and geographic theorical concepts raised by the authors are not only applicable

but essential to acknowledge when devising the operationalisation of the theoretical concepts on the chosen datasets (see section 3.1). For example, the geographical concept of scale is imperative and cannot be ignored, in relation to performing reverse engineering techniques to descramble the preceding point data displacement (Grace et al., 2019, 200). The researchers argue that the current DHS standard for the reassigning of settlement coordinates is inept, mainly due to the geographical characteristics of the settlements, i.e., the chances of two villages on the top of a mountain in the same geographic region (i.e., Tajikistan) sharing similar traits to the point of distinguishable identification, is overlooked by simple coordinates scrambling (Grace et al., 2019, 201). Thus, this issue of environmental homogeneity, results in the potential identification of the study's targeted settlements. The wider question of contextual scale when working with DHS datasets, notably those of a similar nature (e.g., Standard DHS [see 3.3.2]) must carefully when choosing appropriate datasets to analysis in this study (Smith and Mark, 1998, 316-317).

Other fundamentals highlighted by scholars conducting similar studies in this area present a number of broader abstract issues that ought to be considered in the utilisation of this study. In examining the abilities geomasking techniques on similar datasets to the DHS, Seidl et al. (2018; 2015) highlight the importance of research integrity for working with such datasets. In using datasets that apply total randomization of point data, as in the case of the DHS, the data becomes effectively 'reverse proof' for privacy measures. This element does not come without a cost however, most particularly to researchers. While the anonymity of participants and their locations are seemingly protected, the datasets loose some of their veracity (Seidl et al., 2015). By masking the true coordinates in the case of the Tajikistan settlements, using random point displace, in principle, the findings lost some of its true environmental exactness, as emphasized by the authors in their argument highlighting conceptual considerations such as geographical scale (Grace et al., 2019; Seidl et al., 2015). The exposure of such data to a "growing public", particularly those who are not familiar 'map users', can also result in uncertainty of research integrity in applications of these studies in areas outside academic research (Seidl, 2018, p. 475). Seidl et al. (2018, 477) note one of the more serious instances of this happening in the publication of crime maps in the UK by police on their crime statistics website. The maps drawn used geomasking techniques to conceal the true coordinates of affected households and precise clusters of affected areas but failed to illustrate this in such a manner that could be interpretated by a non-GIS operator, thus giving the public a false perception of certain areas in the UK with supposedly high crime rates

(Seidl et al., 2018). This case evidently supports the broader issue that the ever-increasing use of geospatial datasets, by non-geographers is an issue which needs to be addressed. It is on this aspect that the DHS datasets chosen within this study is vindicated (see further discussion on this in section 3.1 and 3.3).

2.4.2 DHS report – How to use spatial covariates

As outlined in the previous section, the DHS produce datasets that contain aspects of demographic, health, and spatial data components (Grace et al., 2019, Dorélien et al., 2013). Prior to 2017, spatial datasets, namely those compatible with GIS software were collected, sorted, and analysed separately to the demographic and health survey data (Trinadh et al., 2018). Researchers, namely geographers, who wanted to use the datasets to conduct a spatial analysis, would be required to manually join GIS datasets, sourced from the DHS's Spatial Data Repository with data from the demographic and health survey datasets, a timely and complicated process (Trinadh et al., 2018). Furthermore, researchers who required additional spatial resources, as seen in the case of Grace et al. (2018)'s use of satellite imagery, would thus be obliged to further merge variables from non-spatial fields (demographic, health, etc) with the externally sourced spatial material. This process can be particularly delimiting when undertaken without a prior knowledge of GIS and spatial analysis, a significant issue previously discussed in 2.5.1 (Trinadh et al., 2018; Seidl et al., 2018). Such was the necessity that the DHS created the Geospatial Covariates, a set of geographic datasets which comprise geographic coordinates that have been recorded by DHS researchers in the field at the precise location of the survey cluster, which are merged with compatible files from the statistical demographic and health survey datasets (Trinadh et al., 2018). This has enabled spatial investigations using the DHS's dataset to occur, without the requirement for geographers to first merge spatial and nonspatial datasets, while also allowing non-spatial researchers such as statisticians and health workers to utilise spatial data such as locational coordinates, without a required foundational knowledge of GIS or use of its platforms (Trinadh et al., 2018).

In addition to geographical coordinates, the spatial covariates contain spatial datasets containing environmental datasets. These range from ecological measurements recorded in the field, to data on climate and environmental determinants (Trinadh et al., 2018). 192 different environmental classifications have been identified and categorised under eight main meta categories: agriculture, climate, environment, health condition, infrastructure, physical earth, political, and population (Trinadh et al., 2018, 2-3). Trinadh et al. (2018, 11, 13) employ a

comparative analysis on environmental determinants (i.e., spatial data) measured against health and other related socioeconomic gauges in the prevalence of childhood anaemia (an illness caused by low haemoglobin in the blood). The study extracts key attributes from the spatial covariates, including demographic surveys, biometric measurements and climate data obtained at the exact coordinates the survey data was recorded (Trinadh et al., 2018).

In briefly discussing Trinadh et al.'s (2018) study which is conducted using the multifaceted spatial, demographic and health Standard DHS datasets, this exemplifies the need for a greater contextualisation of the platform (i.e. the role of metadata, and in particular metadata categorisation) which facilitates geographers to better use and understand multiple dataset types.

Chapter 3 – A study of dataset interoperability using metadata

3.1 Methodological considerations

Considering the main aim of this research is to consider an ontological investigation of an actual data portal's metadata categories, methodological approaches chosen are ones that allow for the best operationalisation of the theorical framework previously presented. This necessitates the merging of theoretical concepts, i.e., framework, with quantitative attributes, i.e., data (in this study particularly, this means data about data, i.e., metadata), hence this study employs a mixed methods approach. The theoretical discussions and concepts outlined in section two form the foundation for the oncolotical examination of metadata categorises. Furthermore, careful consideration is given to the particular spatial datasets that this study examines. In order to do this, a comparative case study approach, guided by the literature, is applied. This facilitates an analysis, both across datasets (i.e., demographic, health and spatial) and across case studies (i.e., the chosen countries, see 3.4). The approach taken to identify, scrutinise and access the datasets used in this study necessitated a suitable and representative data portal. As this research is previously highlighted key debates and integral issues within geoprivacy and geo- ethics in section 2.3 and 2.4, the portal chosen needed to be a closed, verifiable (for both the users and the administrators) platform, preferably run by a reputable, professional organisation. Furthermore, the datasets made available through the portal should be, at a minimum, both spatial and demographic in nature. Any additionally aspects, such as health, add another dimension to the study (see studies by Grace et al., 2018 and Trinadh et al., 2018 in section 2.5). Meeting these requirements ruled out the use of more available, easily accessible, open-source portals such as Open Street Map or the Global Map GitHub.

3.2 The Demographic and Health Surveys (DHS) Program

This section discusses the chosen candidate portal, the Demographic and Health Surveys (DHS), outlining the process involved in sourcing the portal, the main criteria used in its selection and a brief outline of the DHS Program as an organisation and its work.

As noted in the previous section (3.1), the chosen portal has been selected based on certain credentials, namely: its operator and access type, the methods used to collect the data and the characteristics of datasets provided. Preliminary identification of a suitable portal was guided by advice from the thesis supervisor, of whom has expertise and field experience with using datasets from DHS in his own research. Initial scoping of the relevant literature within the field highlighted the DHS as a potential candidate data portal, thus further emphasizing this recommendation (see section 2.4). In their discussion on sourcing relevant data for their own study, Dorélien et al. (2013, 414) highlight the DHS's systematic approach to collection of demographic data, noting that the organisation's constant expansion in its collection methods have allowed for sophisticated spatial datasets (through gathering GPS coordinates and environmental measurements) in addition to the extensive health surveys undertaken to enhance the portal's capabilities in research use. Additionally, Grace et al. (2018) note that the DHS is one of the world's largest and most reputable sources for data on population, health and development which extensively documents and examines the third world.

Established in 1984 by the U.S. Agency for International Development (USAID), the DHS Program provides resources, tools, and researchers to collect household data in aspects of socioeconomic, healthcare, and environmental research, from a local to a national scale, in ninety developing nations. Its operations are utilised through coordination with national governments, NGOs, and research institutions. The DHS operates with the aim of collecting high quality data through sophisticated yet feasible methods for host countries, which can be used as a means of monitoring development, planning, and policy implication. Furthermore, the DHS Program emphasises the host country's right to ownership in data collection, analysis, and use.

The DHS's operational functions of its portal is the main focus in choosing a suitable data portal for this study. Given that the aim of this study is to examine and critique the metadata categorisation techniques employed by the DHS for the operationalisation of its datasets, as contended in section 2.2, the standardisation process is key in facilitating metadata functionality (or interoperability) of multifaceted datasets, thus the candidate portal need to employ an evident standardisation procedure in the processing and classification of its data. Dissecting some of the DHS's key features, the portal has a large library containing a range of different datasets, sourced, and collected through the organisation's 400 different survey types (see 3.3.3). Such data can age as far back as the organisation's establishment in the 1980s, with records from each year updated to be accessible through the DHS's current

storage bank and file formats. The DHS has its own unique classification system for processing, storing, analysing, and presenting data (see 3.3.2), known as the survey process. Significantly, this aspect makes the portal an ideal candidate for testing the research question. In addition to demographic and health surveys, the DHS carries out extensive environmental and spatial research. This material is available through the organisation's Spatial Data Repository, a sub-database of the main demographic and health surveys library. This material consists of the demographic and health survey data provided in a GIS compatible-format (e.g., country boundaries constructed from surveyed areas), modelled surfaces (topography and remote sensing) and population estimates (census records in shapefile format) (Trinadh et al., 2018). In relation to this study, the most important element of the DHS's geographic datasets is its Spatial Covariates. These datasets link survey cluster location (GPS coordinates taken on the survey site, in real time) with secondary data (i.e., the demographic and health survey indicators data) (IFC International, 2013; Trinadh et al., 2018). The spatial covariates datasets are available to access through both the DHS Program main portal (see 3.3.4) and its spatial data repository, with the aim of researchers with limited GIS knowledge (e.g., non-geographers) being able to apply spatial data to their research (Trinadh et al., 2018, 2). Thus, this study, focusing on the DHS demographic and health survey datasets incorporates a fundamental spatial angle through the inclusion of location data.

The final aspect in assessing the DHS's suitability for selection acknowledges the issue of geoprivacy, as highlighted in sections 2.3 and 2.4 (Grace et al., 2018). While the DHS allows free and open access to its presentational material of the datasets (e.g., reports, guides, research publications, etc), access to the raw data (e.g., survey datasets, GPS coordinates, etc) is restricted. This can only be accessed through formal application to the DHS (which is further documented in 3.3) which is reviewed through stages (initial approval, dataset approval, survey-type access, etc). Further, under the guidance of the literature, a brief review of the DHS's ethical framework for protecting the privacy of the survey respondents was undertaken during the preliminary investigative stage. In addition to the measures taken to scramble GPS coordinates and anonymise the locations of the study areas (see 2.4.1), the DHS allows for requests of ethical review documentation for each survey questionnaire and dataset. Participant's identities are anonymised by using a unique serial number identifier on all documents. This step is not essential to this study, but duly noted.

3.3 Study design and steps applied

3.3.1 Applying to the DHS for access to datasets

The first step of the accessing the datasets is to apply directly to the DHS Program through their registration process. While all datasets are free to download and use, gaining access to the portal requires users to set up an account through a registration form. This is done by accessing the portal through the DHS's website (www.dhsprogram.com). Registering for data access is a rigorous process, as access is only granted to what the organisation deems to be for legitimate research purposes. The first part of the registration process is providing the DHS with specific user information. This includes an email address (for verification purposes and contact) phone number, the user's first and last name, the user's institution name (Department of Human Geography, Lund University), the institution type (i.e., university) and the user's country of residence. The second step in the registration process requires the user to complete a concise research proposal (see Appendix 1, Screenshot1). This includes a project title, named co-researchers and their contact details (in this case the thesis supervisor) and a cohesive description of the study which includes the research question, the research design, and a basic data analysis plan. The DHS emphasises at this point that the reviewal is not an automated process. The research proposal will be reviewed by a researcher who may send feedback to the user, including requesting further clarification for some aspects of the research proposal. The third step is to select the datasets required. The selection steps are faceted through categories based on the dataset's region (in this case Sub-Saharan Africa), the countries and the datasets available (e.g., survey, service provision, etc) (see Appendix 1, Screenshot 2). GPS and HIV datasets (the DHS's most sensitive datasets) require additional justification (see Appendix 1, Screenshots 3-5). This process includes a synopsis as to the user's intended use of the sensitive datasets which is separate to the research proposal, and agreement to adhere to the DHS's terms for use and privacy guidelines. Once the request for access is submitted, users must wait for approval (see Appendix 1, Screenshot 6). It should be noted that preliminary requests for standard dataset access takes approximately 24-48 hours to be reviewed, while requests for sensitive datasets (GPS/HIV datasets) can take longer. Further, as the DHS Program is based in America, correspondence from the DHS (regarding dataset requests) often occurred outside of CET. Approval of requests is received through an email containing certified confirmation in writing (attached as a PDF) and instructions for downloading the datasets.

3.3.2 Learning the DHS Survey Process

It is highly recommended, especially for first time users, that a process of familiarisation with the DHS's methodologies for data collection, processing and publication be undertaken by the user, prior to requesting access for datasets. Furthermore, this allows the user to be affective and compliant in utilising the data in their own research. The pre-selection of datasets to analyse first requires knowledge of the DHS's surveys: by country, by year, by survey type (e.g. Standard DHS, SPA, etc) and by survey characteristics. Additionally, reviewal of the corresponding publications (info-graphs, final reports, survey questionnaires, etc) of the datasets is a key step in gaining an understanding for both the DHS's methodological approaches and the selected study area(s). This also aids in developing the study design for this research, as subsequent documentation such as survey questionnaires act as a template guide for setting out research analysis requirements. It can be noted, this step in the DHS's process clearly suggests that an emphasis on metadata and categorisation is significant (see 4.2 for further discussion).

The key aim of this study in reviewing the DHS's survey process is to determine how the datasets are constructed and thus, determining the lifecycle of the datasets. This starts with studying the *survey process* framework, an outline of the steps taken to gather data within the field by DHS researchers. This process consists of four major steps and takes approximately two and a half years to complete, as illustrated in figure 2 (The DHS, n.d.b; The DHS, n.d.c).

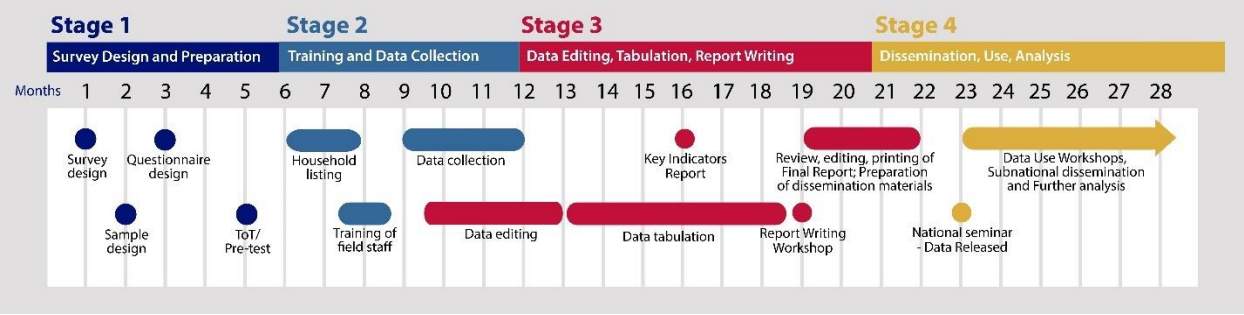


Figure 2: The DHS Survey Process, taking up to 2.5 years to complete.

Source: The DHS (n.d.c)

Stage one is designing the surveys: the questionnaires, sampling manuals and nutritional/health-related briefs, which takes approximately six months. Stage two of the survey process focuses on field work and personnel: training interviewers and research staff in tasks such as mapping, GPS data collection and biomarkers. In addition to data collection,

researchers become accustomed to the local cultures, customs, and environments. Field teams for interviewing and health measurements are divided according to their role. A notable example includes gender separation: female researchers usually head interviewing tasks, particularly where the head of a respondent household is female or when the survey topic is deemed sensitive (e.g., sexual health, domestic violence, etc) (The DHS, n.d.**b**). This element is important to note when reiterating the arguments made by Critical GIS scholars such as Pavlovskaya (2018) and Elwood et al. (2011), who emphasize elements such as positionality, epistemological background, and the utilisation of GI technologies for positive social change (see section 2.1). Stage three involves data editing, data tabulation and report writing (i.e., the formation of the DHS's own meta categorisation for datasets). This stage is started in conjunction with stage two, however can take over a year to complete in full. Stage four concentrates on dissemination of the datasets and their publications. Taking at least six months to complete, this process includes research seminars, peer review and further analysis, typically in concurrence with the host nation's governments, NGOs, and research volunteers. In reviewing the DHS's survey process, it can be determined that a proprietary data lifecycle is present. In reiterating the objective of this research, to examine the metadata categorisation used by the DHS on its datasets and how this can facilities the dataset's users, the intimal steps within the DHS's survey process further demonstrate the emergence of a standardisation procedure, a vital step in the formulation of, and most importantly, operationalisation of metadata functionality (see 2.2 and 4.2). The final process in determining the lifecycle of the data is to learn about the DHS's own survey types. While the DHS holds over 400 different surveys, these are categorised into five main classifications: Demographic and Health Surveys (DHS), Malaria Indicator Surveys (MIS), Service Provision Assessment (SPA) Surveys, [Other] Quantitative Surveys and Qualitative Research. These are further sorted based on country, year, and characteristics (i.e., survey topic, mainly health, e.g., pregnancy history, tobacco use, etc). For this study, the Demographic and Health Surveys (DHS) has been chosen as the candidate survey type (see section 3.3.3). As outlined previously, stage two of the data collection process sees the collection of demographic data, in addition to GPS data collection and biomarkers. Biomarkers are biological measurements of health conditions. Differing from survey characteristics (qualitative health data collected through questionnaires and interviews), biomarkers are quantitative measurements of a person's physical health (e.g., height and weight, blood pressure, etc). The biomarkers data are stored in numeric format and categorised based on measurement of nutrition, health, immunization, STIs and environmental determinants.

3.3.3 Choosing candidate survey datasets

As outlined in 3.3.1, initial interaction with the gatekeeper of the data involves a lengthy authentication process in order to gain access to the datasets. Prior to approval, the DHS provides model datasets and questionnaires free to download without preapproval, with the intention of allowing users to familiarise themselves with the file layouts and practice using the datasets. Utilisation of practice datasets further provides guidance on the types of metadata categorisation used by the DHS, which is discussed further in Chapter 4.

When an access request is approved (see Appendix 1, Screenshot 7), the user enters the Download Approved Datasets page which shows the datasets that they are eligible to download. Initial survey data for Tanzania was requested and approved, including datasets for all available years. Additional requests for sensitive GPS and SPA (both spatial) datasets were approved (see Appendix 1, Screenshot 8). HIV datasets were not approved for use in this study as justification did not meet the research criteria. Approved datasets are categorised based on year, survey type (sensitive datasets are sorted separately from standard, see Appendix 1, Screenshot 7). and phase (see 3.3.4). As stated in 3.3.2, the Demographic and Health Surveys (DHS) is the focus for this study. The DHS, or Standard DHS, is the typical survey type carried out by the organisation and includes demographic, health surveys and biomarkers, which deemed it a suitable candidate for the purpose of this investigation.

The next step involves preparation for download. Opening the Standard DHS 2015-16 (the most current dataset) takes the user into the *survey dataset file* directory (see Figure 3, in 4.2 for illustration), a separate file portal which stores all the datasets available to download in different file formats (see 3.3.5). Each file is categorised using variables, known as the DHS Recode Data. This is the name used for the data processing procedure, transforming the original raw data (e.g., paper-based interview scripts, questionnaires, field notes, etc) into a standardised format (ICF, 2018, 4). The recode allows datasets to be further sub-categorised (e.g., all survey datasets specific to males is filed under the metatag *Men's Recode*) (see Appendix 1, Screenshot 8). In addition to this, the corresponding geographic datasets (GIS and Spatial Covariates) and biomarkers for the Standard DHS 2015-16 surveys are sorted in this portal (see Appendix 1, Screenshot 8).

A second request was later made for access to datasets for Uganda, including additional requests for GPS and SPA datasets, but not HIV (see Appendix 1, Screenshot

10,12). Notably, the Uganda datasets are marked by the DHS in the data selection menu as including restricted datasets (see Appendix 1, Screenshot 11). Upon receiving user confirmation for request for the Uganda datasets, the DHS includes notes on restricted access to certain datasets [survey: AIS 2004-2005] (see Appendix 1, Screenshot 13). A search through the DHS's report catalogues later revealed the restricted datasets to be a behavioural study for HIV/AIDs, containing extremely sensitive records, the raw data of which, once standardised into variables, was destroyed for privacy measures (see 3.3.5) (Ministry of Health (MOH) [Uganda] and ORC Macro, 2006). Access for the remaining Uganda datasets was subsequently approved (see Appendix 1, Screenshot 14).

3.3.4 Navigating and considering the DHS Metadata Categories

As noted in previous sections, the DHS follows a strict structured process in their data collection, processing, categorisation, and analysis. This section focuses acutely on the categorisation procedures, with the aim of highlighting potential MDCs to be evaluated in this study's ontological examination. As stated previously, datasets that have been approved for download are sorted based on year and survey type. The additional label of *phase* is also used as part of the description for each dataset (see Appendix 1, Screenshot 7, 14). Data collection for the Standard DHS surveys is carried out every five (the number of years differs for other survey types), with each phase differing from the previous (e.g., the inclusion of the spatial covariates with more recent datasets). To date there are seven (labelled *I-IIV*) phases. Users should be cognisant of the changes made in each phase of collection, particularly when analysing more than one of the same survey types from different years (see 4.2 for further discussion).

The use of metadata categorisation by the DHS is also observed at the sublevel. Individual datasets available to download, e.g., the Men's Recode for Tanzania [TZMR7B.SAV], are categorised within their assigned variables (see Appendix 2, dataset citation). The DHS refer to this as the file *naming convention*. Using dataset file TZMR7B.SAV as an example, each set of characters within the file name are a subcategory. The first pair of characters, ("TR") refer to the country code (Tanzania). The ("MR") characters refer to the dataset type, namely, the recodification label for the surveys concerned (Men's Recode). This classification is also important when referring to the data in research, as it is used as the unit of analysis for datasets. The next characters ("7B") refer to the release version of the survey (e.g., if it is the first survey of its type conducted during a particular

phase). The remaining characters (".SAV") refer to the file format (an SV file to be used in SPSS).

Other sublevels of meta categorisation used by the DHS includes meta tags on dataset's characteristics, including but not limited to: the implanting organisation (what governing body commissioned the study), highlighted research topics (see 3.4 for case-specific examples) and metadata for respondents (sample sizes, etc) (See Appendix 1, Screenshot 9).

The levels of categorisation and sub-categorisation observed also re-affirm the lifecycle of the datasets principle outlined in 3.3.2. Further, this task aids in the selection of contrasting datasets (i.e., Tanzania and Uganda) for the purpose of a comparative analysis (see 3.4). Further analysis and empirical discussion of the DHS's use of distinguished meta fields, metatags and unique categorisation methods is presented in at length in Chapter 4.

3.3.5 Extracting datasets files

The final step is to download the chosen datasets. This procedure is significantly aided by the DHS's "download manger" program, which facilitates the download of both individual or multiple datasets at once. All survey datasets are stored as compressed .ZIP files as they contain multiple files which are opened using different platforms. The survey and biomarkers datasets are formatted as flat data (.FL), SPSS data (.SV), Stata data (.DT) and SAS data (.SD). Geographic datasets are formatted as shapefiles (.SHP) for GIS data and an Excel spreadsheet (.csv) for the Geospatial Covariates. The appropriate application packages (e.g., ArcGIS, SPSS, etc) are required to be installed in order to open the dataset files once downloaded.

Subsequent the download process, multiple dataset files can be joined or merged, depending on the file format. As previously mentioned, the geographic and biomarkers datasets provided (upon permission granted from the gatekeeper of the datasets) are designed to supplement the survey datasets. Consequently, the action of combining the datasets (demographic, health and spatial) could increase the potential risk of disclosure of participant's private information, given the interoperable nature of the datasets. While this aspect is not a core focus of this study, further discussion on how this particular feature of

metadata functionality can in turn be as much of a concern, as it is valuable to researchers, is outlined in section 4.2.

3.4 Background and context of selected candidate datasets

The Sub-Saharan nation of Tanzania is the first candidate country analysed in this study. Tanzania was chosen as the initial area of study, led by guidance from the research supervisor. His own research uses Tanzania's DHS datasets and from this, recommendation was made that the Standard DHS surveys from Tanzania would be an ideal dataset for a beginner and an ample reflection of a full DHS dataset. Further, his expertise could aid in flagging potential issues or areas of concern that could later be addressed when discussing the metadata categorisation of the datasets on a micro-level (i.e., case or country specific issues). The 2015-2016 Standard DHS survey was collected in the seventh DHS phase and examined a representative sample of around 17,000 participants (see Appendix 2, Report 1). Among the standard areas of research surveyed, the Tanzania Standard DHS focuses on two particular areas of priority: Malaria (prevalence, participant knowledge and prevention) and women's issues (empowerment, FGM and domestic violence (see Appendix 2, Report 2).

The second candidate country, Uganda, was chosen through investigation on data poverty and access rights in Africa. Evidently, the World Development report commissioned by the World Bank (2021, 58) focusing on the data rights of in impoverished nations, reveals that data rights watchdogs, namely, civil society organisations (CSOs) in Uganda demanded that all contracts with private entities, where public procurement projects are concerned, be open, freely, and easily accessed by its CSOs, on behalf of the general public, through monitoring of local administrative databases. This call for transparency and social accountability of a previously corrupt area in governance thus made Uganda the first and only African nation to enact change in its national procurement standards, by demanding data access for the public (World Bank, 2021; GPSA, 2020). Since the area of public procurement ties in closely with the work of the DHS, consideration was given as to the nature of the potential datasets that could be available, given the noted indication of civic engagement with data rights in Uganda as suggested in the World Bank's report. The 2015-2016 Standard DHS survey for Uganda was collected in the seventh DHS phase, with a representative sample of around 23,000 participants (see Appendix 2, Report 3). Among the priority research topics

include: Nutrition and HIV (knowledge, attitudes, and behaviours) (see Appendix 2, Report 4).

It should be noted that both datasets are among several featured and analysed by Trinadh et al. (2018), as outlined in section 2.5.2. Examining the application of the datasets in a DHS-authorised report significantly strengthened the justification in selecting both datasets, having observed both fully analysed and applied to a model study.

3.5 Study summary

In summary, it is clear having carried out this study thus far, that the DHS use metadata as a guiding mechanism both within and across datasets, and across studies (i.e., a common meta format for all countries, regardless of the raw data content, e.g., Tanzania and Uganda). Furthermore, the DHS portal itself is built based on a process model, its Survey Process, which in turn acts as a means of standardising its data collection methods, from early field fieldwork through to the publication of datasets and their results report. In addition to this, it is evident that without a deeper understanding and access to the contextual knowledge systems which oversee these processes, the more difficult it will be for users, namely first-time users, to access, understand and ultimately utilise the datasets correctly. It should also be noted that the time and commitment required to fully learn and utilise this process is considerable, with the additional constraint of having little prior knowledge within the other field areas (i.e., health data, statistical analysis, etc). Finally, to conclude on a positive note, it is clearly apparent having accessed and infiltrated the datasets, through to their sub-level categories, that the degree of interoperability between the different dataset types (i.e., spatial, demographic and biomarkers) is significantly high, given their common meta-architecture and the number of different file types available.

Chapter 4 – An ontological approach to Metadata

4.1 Metadata as a guiding mechanism

This chapter presents the outcomes of conducting the study. In doing so, it examines the DHS's metadata categorisation in two ways: firstly, on a semantic level, i.e., the role of the metadata categories and how they aid and assist the user, and secondly using an ontological lens, i.e., delving deeper into the DHS's specific taxonomies and specialist metadata layers, examining how effectively they capture the complex real world and tacit phenomena encountered by researchers in the conduction of the DHS Survey Process. Furthermore, this section examines the degree of interoperability that can be implemented while further assessing the levels of ontological approach can be applied to reviewing and evaluating the DHS's metadata. Before discussing the outcomes of this study, this section briefly reexams the nature of metadata specifically against the datasets employed in the study presented in Chapter 3.

The principal role of metadata is to act as a bridge between the highly conceptual knowledge systems employed by researchers and the obstinate analytical data semantics used within software (Schuurman 2009b; Schuurman and Leszczynski, 2006). Taking Smith and Mark's (1998, 308) observation, purporting "the nature of being", metadata has a significant impact on the interpretation of data values, bringing a deeper contextual understanding for dataset users by providing data with interconnected layers of meaning and thus allowing geographers to better navigate and cross-compare the multifaceted layers of data, such as that of the Standard DHS datasets. Elwood et al. (2011) consider how a researcher's epistemological background can also impact the interpretation of metadata categories, and thus, contextualisation of datasets. Furthermore, Schuurman (2005) illustrates this point with relevance to geographers through her 'object' discussion: what one spatial data user may rationally view as point data on a map, metadata categorisation allows another to interpret said data as a house, someone's home, thus bringing contextual dimensions to the factual statement expressed by the spatial data points. Guided by these concepts outlined in the Theoretical Framework, this study observed that the Demographic, Health and Spatial datasets within the Standard DHS files overlap, as illustrated in figure 3.

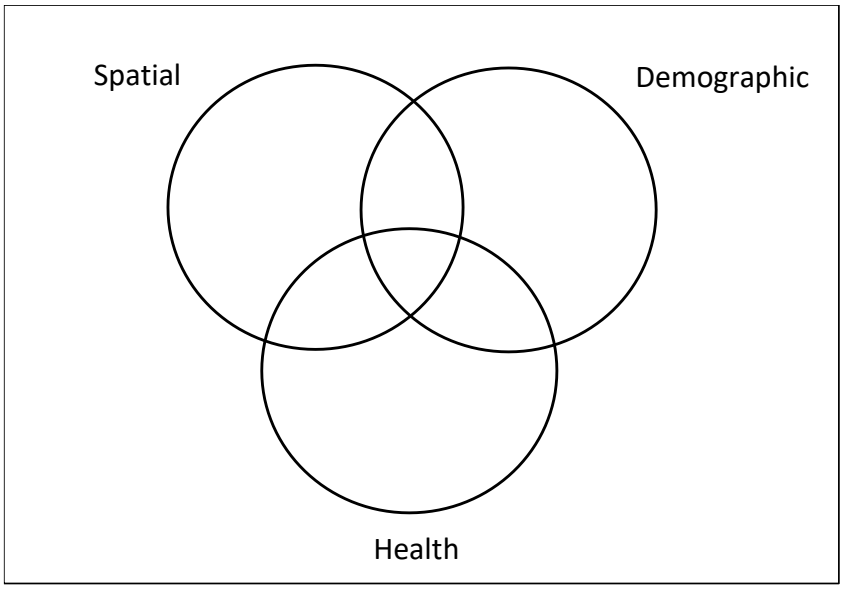


Figure 3: Overlapping DHS dataset types.

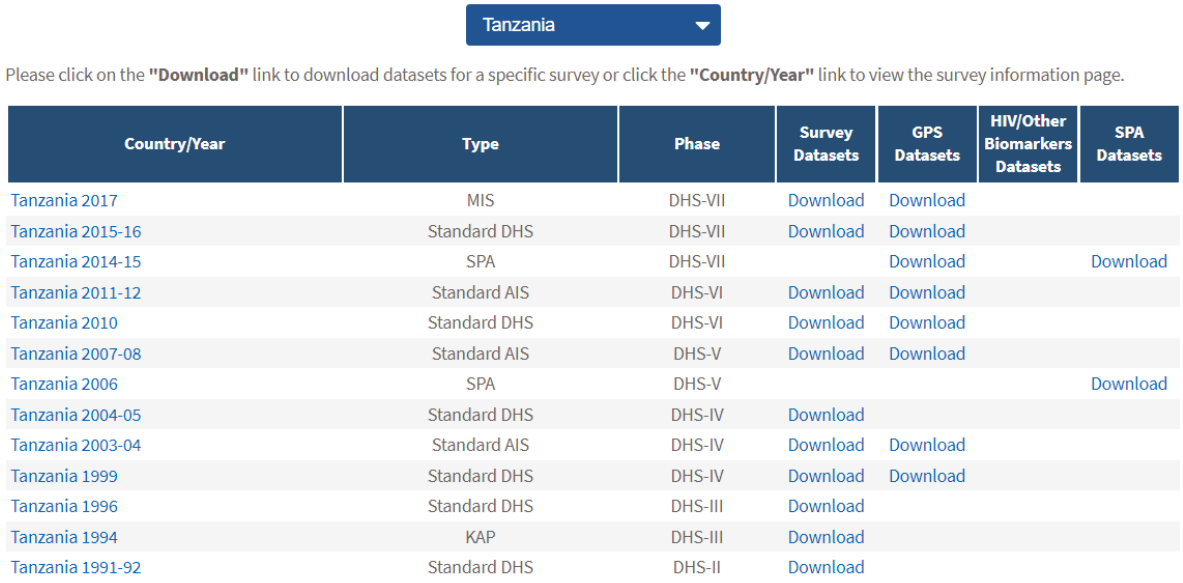
In examination of the DHS portal, this study finds that while the DHS is an ideal platform for allowing geographers access to multifaceted datasets, i.e., spatial and non-spatial, utilisation of said datasets (i.e., interoperability), however, can be hindered without a logical, but contextual, guiding framework. Thus, the role of metadata categorisation in this process becomes the key in enabling geographers to understand and ultimately utilise the datasets. Affirming this position, Schuurman (2009b) argues that metadata has the ability to "… become a repository of qualitative information about both quantitative and qualitative spatial and non-spatial attributes. In the process, metadata will enhance the ability of GIS users to incorporate multiple knowledge systems – or ontologies". In undertaking a deep probe of the DHS's operational process (through examining the lifecycle of the data), this study observed the datasets to be multi-dimensional (e.g., all datasets regardless of the type included a spatial element, e.g., corresponding geographic coordinates). Therefore, the wider availability of metadata to guide users through the process, the more geographers and other spatial users can infiltrate and gain a deeper understanding of the non-spatial datasets available, leading to the potential inclusion of datasets such as demographic and health-markers within GIS and spatial research outputs. This reaffirms the role of metadata, expanding the "limited concept" that its sole purpose is that of data documentation and ultimately facilitating the inclusion of "multiple knowledge systems" through interoperability of multifaceted datasets (Schuurman and Leszczynski, 2006; Danko, 2008; Schuurman, 2009b).

4.2 Understanding context and meaning using MDCs and MDTs

On examining the formulation of certain metadata fields and categories within a standardised system, Schuurman and Leszczynski, (2006, 717) stress that "[i]n order to understand the context in which an attribute is used, it is important to understand the rationale for its collection". The first part of this analysis examines how the DHS's metadata categorisation (MDCs) and tags (MDTs) aid users on a surface level, in navigating its vast file directory. The second part focuses on the DHS's specific taxonomies, conducting a deeper contextual analysis, evaluating how this can facilitate users, namely geographers, in better utilisation of the datasets.

The DHS uses its own proprietary model in the creation, storage, and purpose of its datasets. As outlined in section 3.3.4 in Chapter 3, the DHS portal is structured using several file directories made up of complex layers which contain the datasets. Datasets are sorted and compartmentalised using an intricate metadata categorisation system which derives from the DHS Survey Process (see 3.3.2). As noted in section 2.2, metadata typically comes in a statistical text file format, available to read through online catalogues or through programs such as ArcCatalog (Danko, 2008; Schuurman and Leszczynski, 2006). From a GIS perspective, metadata facilitates interoperability through standardisation (Schuurman and Leszczynski, 2006). An immediate observation upon gaining access to the dataset file directories for both Tanzania and Uganda (see Appendix 1, Screenshot 7, 14) is the uniformity of how the data is sorted and stored. As highlighted in 3.4, each case (or country) presents different research topical outcomes, determined by the nature of the data collected in the field. This, however, is not apparent, on a surface level, specifically due to the standardisation of all metadata categorisation and semantic nature of the fields and tagging for each country and dataset available. While this initial standardisation process makes the datasets more uniform in an operational capacity (i.e., interoperable across datasets, platforms, software packages, etc), a deeper contextual understanding as to why the metadata tags and categories are formulated the way they are, is required by the user. As argued by Schuurman and Leszczynski (2006, 716), metadata categorisation allows for a better evaluation of the datasets, most notably, how they can be used by geographers. To achieve this however, metadata needs to have: context (of the data collected, i.e., the semantic approach) and a mandate (i.e., how to use the data and what for, i.e., the ontological approach). Examining the semantic, or surface level, roles of MDCs and MDTs, figure 4

illustrates a typical DHS survey dataset file directory, in this case available for the country Tanzania. Illustrated in the blue bar at the top, the metadata categories, or MDCs formulated by the DHS, provide the user with a logical, coherent classification structure for accessing the datasets, sorted by year, survey type, and dataset sorts (e.g., demographic, health, and spatial).



Country/Year	Type	Phase	Survey Datasets	GPS Datasets	HIV/Other Biomarkers Datasets	SPA Datasets
Tanzania 2017	MIS	DHS-VII	Download	Download		
Tanzania 2015-16	Standard DHS	DHS-VII	Download	Download		
Tanzania 2014-15	SPA	DHS-VII		Download		Download
Tanzania 2011-12	Standard AIS	DHS-VI	Download	Download		
Tanzania 2010	Standard DHS	DHS-VI	Download	Download		
Tanzania 2007-08	Standard AIS	DHS-V	Download	Download		
Tanzania 2006	SPA	DHS-V				Download
Tanzania 2004-05	Standard DHS	DHS-IV	Download			
Tanzania 2003-04	Standard AIS	DHS-IV	Download	Download		
Tanzania 1999	Standard DHS	DHS-IV	Download	Download		
Tanzania 1996	Standard DHS	DHS-III	Download			
Tanzania 1994	KAP	DHS-III	Download			
Tanzania 1991-92	Standard DHS	DHS-II	Download			

Figure 4.

Source: Appendix 1, Screenshot 7, Tanzania dataset.

Figure 5 is an illustration of what the file directory would look like without the use of top-level MDCs. Note that the individual datasets are still labelled using metadata tagging, a sub-level of labelling. While categorisation is the broader classification, tags or MDTs provide a concise contextual descriptor for the dataset. The contrast between the two figures illustrates the necessity of the hierarchical metadata categorisation process that the DHS impose, while also demonstrating the importance of sub-level tagging and labels, particularly, in this case, for datasets specific to the DHS's own methodological practices, such as phases (see 3.3.4) and survey type (see 3.3.2). Additionally, this process also allows for the complex metadata classification illustrated, to be converted semantically into a short, readable text format, using concise MDCs and MDTs labelling (see 3.3.4).

Tanzania 2017	MIS	DHS-VII	Download	Download		
Tanzania 2015-16	Standard DHS	DHS-VII	Download	Download		
Tanzania 2014-15	SPA	DHS-VII		Download		Download
Tanzania 2011-12	Standard AIS	DHS-VI	Download	Download		
Tanzania 2010	Standard DHS	DHS-VI	Download	Download		
Tanzania 2007-08	Standard AIS	DHS-V	Download	Download		
Tanzania 2006	SPA	DHS-V				Download
Tanzania 2004-05	Standard DHS	DHS-IV	Download			
Tanzania 2003-04	Standard AIS	DHS-IV	Download	Download		
Tanzania 1999	Standard DHS	DHS-IV	Download	Download		
Tanzania 1996	Standard DHS	DHS-III	Download			
Tanzania 1994	KAP	DHS-III	Download			
Tanzania 1991-92	Standard DHS	DHS-II	Download			

Figure 5: The Metadata Categories removed.

Source: Appendix 1, Screenshot 7, Tanzania dataset.

Furthermore, from an ontological perspective, the lower-level MDTs provide spatial users with context (i.e., description of the data) while the top-level MDCs offer a broader mandate for the data (i.e., dataset utilisation in accordance with how it has been formulated) (Schuurman and Leszczynski, 2006, 716-717). Conclusively, the use of MDCs and MDTS on a surface level allow the user to be guided through the dataset portal, with a limited but sufficient understanding of the dataset's configuration.

In stating this, the analysis now moves from examining the semantic approach to metadata, to an ontological approach to metadata, specifically looking at the DHS's individual metadata categories, tags, and processes of standardisation. It focuses on the DHS's distinct MDCs of which this study found to merit further analysis and examine the rationale behind their formulation. This is guided by the framework devised by Schuurman and Leszczynski (2006, 718) in their study of capturing metadata fields ontologically (see concepts: 'sampling methodologies' and 'collection rationale' outlined in Figure 1, 2.2) and other theoretical concepts outlined in Chapter 2. As previously stressed in section 3.3.2, the DHS insist that all users, regardless of research background or intended use of the datasets, study and learn about their methodological practices, namely, the DHS *Survey Process*. Two key observations are evident in analysing this procedure. First, the DHS Program is meticulous throughout its collection process, from fieldwork to the dataset publication, with the intent of metadata categories in the portal being formulised to reflect the stages of the survey process (e.g., year, phase, survey type, etc). Secondly and most notably, the DHS Survey Process itself provides a conventional architecture for standardising the raw data collection and processing, and thus,

its metadata categorisation. Reaffirming Danko (2008) and Percivall's (2008) commentary on the formal standardisation practice across all spatial platforms and dataset types for metadata, the DHS, through its survey process, follows the same procedure as the ISO/TC and other governing bodies, by employing a common meta prototype across all its datasets. Further, unlike most spatial meta frameworks, learning the process prior to utilising the datasets provides a deeper contextual meaning behind the data, thus facilitating enhanced interoperability across multifaced datasets (Danko, 2008; Schuurman and Leszczynski, 2006). Ensuing this, the DHS's contextually layered taxonomies also provide users with a greater knowledge for implementing the datasets. The first category of significance is the *Phase* classification. On a surface level, DHS phases refer to a specific time period between conducting each survey and data collection (typically 5 years). However, applying a broader analysis to this MDC shows that phasing involves a constant upgrading and enhancement of collection methods, survey deployment, field methods and data cleaning and processing. This in turn can have a significant impact on the raw data files which lay beneath the metadata tags. A notable example of this is the inclusion of geographical coordinates to accompany all dataset types collected in the latest phase (VII), using real-time datapoints captured in the field, meaning all datasets hold a spatial functionality, regardless of their primary description. Thus, this reiterates the necessity for DHS portal users to exercise a deeper ontological understanding of the rationale that is behind the MDCs construction. Schuurman and Leszczynski (2006, 718) note this as one of the key elements when ontologically assessing metadata, under their concept 'Data model' (see section 2.2, figure 1). The next category of significance to this study is the DHS *Recode Data* categorisation. As previously outlined in 3.3.3, individual files available to download within a dataset is categorised using an assigned variable (see 3.3.4 for a detailed breakdown of each variable character meaning). While the characterisation, or file naming convention, is a concise, convenient method to sort and categorise the data variables semantically, this method demonstrates a clear example of why further contextual elaboration is required by users when navigating the datasets. Schuurman and Leszczynski (2006, 718) also emphasize this element within their framework, under 'definition of variable terms' (see 2.2, figure 1). Further evaluation, using the DHS Recode Manual, reveals the primary reasoning behind the recode data standardisation process, some, but not all of which is related to data cleaning and data normalisation for statistical analysis (ICF, 2018, 1). Other motives simply cannot be captured semantically and instead are required to be further evaluated by researchers (ICF, 2018, 1). These include respondent data related to imputed dates (e.g., birth date, last use of contraception, etc) (ICF, 2018, 1-2). Other

variables collected in the raw data may serve as a convenience for data collection but not for analysis (e.g., sensitive topics within women's health such as fertility preference questions, i.e., the number of planned children, pregnancy issues, etc). Accompanying each dataset is also the individual survey's questionnaire booklet, which can include user recommendations and notes from the field. The DHS recommends users to study the format of the questionnaire when examining variables, as this can provide further rationale as to country-specific variables (e.g., the restriction placed on the dataset for the Uganda AIDs survey [AIS 2004-2005], identified and discussed in section 3.3.3) (ICF, 2018, 2). Schuurman and Leszczynski (2006, 717) also advise similar practice, under the 'policy constraints' and 'anecdotes' fields within their framework (see figure 1, section 2.2). While the primary result of this study has found metadata to be the guiding pathway for spatial users in operating the DHS portal, this illustration further justifies the prerequisite for an ontological-based approach to metadata, as established semantical practices can't always provide complete context to the dataset structure or its formulation. More simply, the complex real-world occurrences and phenomena cannot always be captured in a logical, connotational system. Finally, the most significant outcome within this section, the DHS, as the gatekeeper of the data, prioritises interoperability across datasets for its users. Noted as being "… in many ways most important[ly]", the data recodification process seeks to standardise dataset formats through its data variables names and coding categories (the metadata fields) so as to facilitate an "easy comparison of data between countries" and across dataset types (ICF, 2018, 2). Moreover, this study finds that in conducing an ontological review of the DHS's MDCs and MDTs, the DHS enables its users, e.g., geographers, to amply utilise its datasets through its sophisticated yet artfully instructive dataset taxonomies.

4.3 Managing interoperability through metadata

4.3.1 Interoperability between datasets

The Encyclopaedia of GIS describes interoperability as the ability to merge several types of datasets (e.g., demographic, health and spatial) and further, utilise them using the same file or program formats (Danko, 2008). As discussed within the Critical GIS framework in 2.1, GI and other spatial technologies have moved toward a more progressive research ethos, i.e., incorporating, rather than excluding different knowledge systems (Elwood et al.,

2011; Pavloyskaya, 2009; Schuurman, 2000). Focusing on geographers, debates within Critical GIS observe the epistemological clashes that occur between the sub-disciplines and schools of thought (Elwood et al., 2011; Schuurman, 2009a). This in turn means that research portals such as the DHS need to be versatile, not only in the type of data they capture, but also in how it can be utilised. Focusing on geographers, the DHS provides various types of data within the broader category of 'spatial' datasets. These include population data, environmental and climate data, and GIS and locational data (see 2.4.2). Thus, the DHS operates through multiple levels, or degrees of interoperability, by providing more than one kind of dataset within a fixed classification type. In addition to this, the broader dataset types (demographic, health and spatial) are designed to overlap, (see figure 3, 4.1) and most importantly be interoperable (see 4.2).

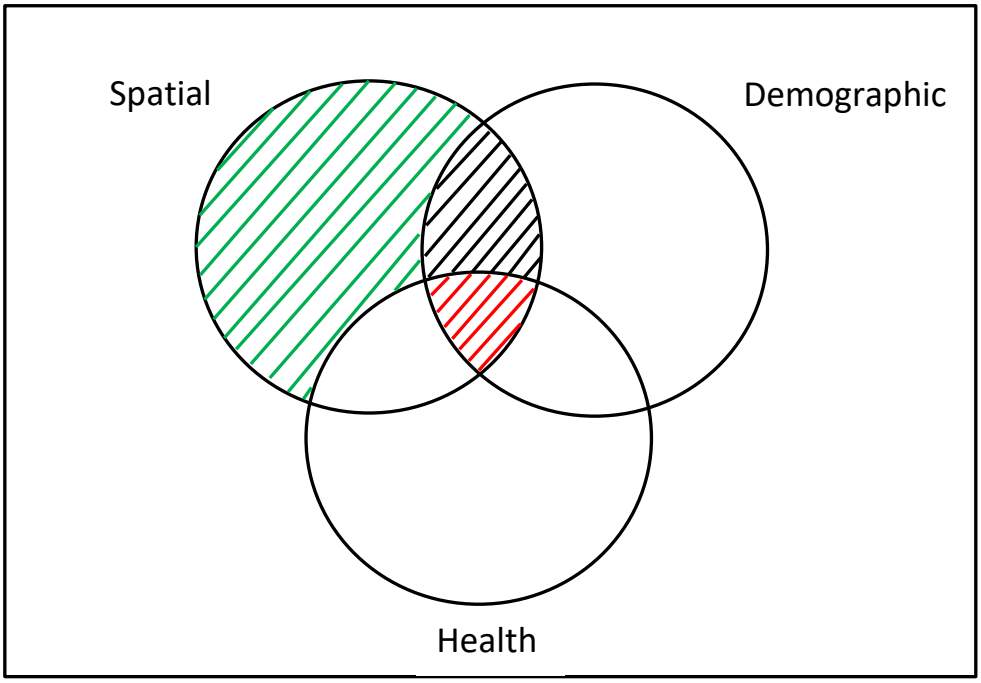


Figure 6: The DHS dataset types operate through various levels of interoperability.

The diagram above (see figure 6) illustrates the multiple levels of interoperability that DHS datasets are designed to function within. Using a colour graphic to represent the degrees of interoperability, green illustrates an individual dataset type (e.g., spatial), which includes the sub-categories of dataset types that can be employed, as previously discussed. Black represents the intersection between more than one dataset type (e.g., spatial and demographic) with the DHS designing datasets to enable metadata functionality through the common metadata categorisation and common file structures (see 3.3.5). Furthermore, the red intersect

illustrates ultimate interoperability, i.e., the joining and utilisation of multiple dataset types (e.g., spatial, and demographic and health), facilitated through a conventional standardisation procedure (i.e., the DHS survey process) and multiple file types (see 3.3.5) for utilisation. While enhanced interoperability across datasets is fundamental for researchers, it is important to note in line with the concepts discussed in section 2.3 and 2.4.1, that continued intersection of datasets and its data, can in turn, increase the potential risk of disclosure of the participant's information. The merging of datasets, across interoperable platforms ought to be handled carefully, given the sensitivity of many of the dataset's content. Furthermore, researchers, such as Grace et al. (2019) and Seidl (2018; 2015) have demonstrated that even the most sophisticated techniques such as geomasking can be compromised (see 2.4.1) and thus, reemphasise the importance of evolving "good practice", i.e., ensuring continued development of geo-ethical frameworks in line with the development of new spatial technologies (Swanlund and Schuurman, 2019; Longley et al., 2015; Rambaldi et al., 2006). In illustrating this process, this study has two key findings from this outcome: the DHS portal facilitates interoperability on a broader level, across datasets, as illustrated in figure 6. Furthermore, in recognising the needs of its users, i.e., the research community, the DHS also provides interoperability within a single dataset type (e.g., spatial), by providing a variety of data types which reflect the diversity of research methods within the field (e.g., climatology, environmentalism, and GIS fields within Geography).

4.3.2 An ontological approach to metadata

As established in section 4.1, metadata provides the missing link between data platforms, by constructing a skeleton across datasets, e.g., Tanzania and Uganda, spatial and non-spatial, etc (Schuurman 2009b; Schuurman and Leszczynski, 2006). As seen in 4.2, most notably with the data recode procedure, practicing exegetics on data, i.e., attempting to translate abstract, messy social phenomena into a workable semantic code, can be difficult, thus a standardisation process is required, most especially to facilitate interoperability. As discussed in 2.2, standardisation is achieved through research consensus which involves incorporating multiple knowledge systems and providing contextual understanding to the attributes and characteristics which make up the structure of the system (Danko, 2008; Percivall, 2008; Schuurman and Leszczynski, 2006). Building upon this, the discussion now moves towards the study's key outcomes. In conducting an ontological approach to metadata,

this study has analysed the DHS's metadata categorisation system, by examining the DHS survey process and utilising two of its datasets (Tanzania and Uganda). This study finds that an ontological approach can be applied to evaluating metadata in three levels. The first level of application is to review, familiarise and understand the metadata categorisation system, on a surface level. As discussed in the first part of section 4.2, metadata tagging and broad categories provide the user with a guiding pathway through the DHS portal's complex, multi-layered file directory. As referred to by Schuurman and Leszczynski (2006, 211) as the data "dictionary definition", these semantic terms allow a user to review the datasets which they wish to utilise, while providing a basic understanding of the dataset's context based on the sorting and categorisation, more so than the data's actual content. Furthermore, reviewing the DHS's user operation manuals and grasping the DHS survey process will further contextualise the metadata categorisation and facilitate utilisation. The second level of application encapsulates the definition of ontology, "the nature of being". This involves the user interconnecting the datasets with their background context. As outlined in 4.2, a deeper, more comprehensive analysis of the DHS survey process and field practices, provides the user with collection rationale, breakdown of the classification system, definition of variable terms (naming convention) and the most importantly, anecdotes (as defined by Schuurman and Leszczynski, 2006, 717, "additional comments pertinent to understanding how to use the database") (see figure 1, section 2.2). This arms the user with as much contextual information as possible, of which may not be possible to include in the initial semantic make up of metadata categories. By the nature of being, a deeper contextualisation of the background actors and processes which provide input to the dataset makeup, allows the user to further factor in the complex human activities which cannot be captured semantically, into their analysis. Additionally, this process further enables the user to implement the datasets with a degree of interoperability, particularly if the user comes from a particular research background (in this case Human Geography, thus spatial). Metadata functionality allows the user to operate across dataset types and specific cases (e.g., country). The third level in implementing an ontological approach in this study, entails the further taxonomizing of the metadata categories. As outlined in 4.3.1, the DHS employs various levels of interoperability within datasets (e.g., spatial) in addition to across datasets. It is important for the user to recognize that within MDCs, lay various degrees of sub-categories, that may be country-specific (as seen with Uganda, outlined in 4.2) or differentiate significantly (e.g., Phases). Cynically, Schuurman (2009b) affirms that the low utilization of metadata is imputable to low completion rates of metadata forms by data producers and further, the limited nature of said

metadata provided. The DHS portal defies this assumption by providing the user with multifaceted categorisation and the complete picture of the lifecycle of a dataset, through its survey process. Thus, in turn, the user, namely geographers, have the capacity to engage with its metadata directly and utilise it both semantically, i.e., in a catalogue or documentational function, and, ontologically, i.e., implement the meta fields, layers and taxonomies in a qualitative capacity, using it to gain further understanding and contextual description of the dataset's content and collection rationale. Further, this study finds that both outcomes, but most notably the latter, will significantly aid in enhanced interoperability.

Chapter 5 – Findings

5.1 Findings

This section discusses the key findings from the conducting of the investigation and the results of the study. In doing so, the research aimed to establish primarily, how can an ontological approach to metadata facilitate interoperability across spatial and non-spatial datasets, while subsequently questioning what benefits geographers could gain by employing an ontological approach alongside established semantic approaches to metadata. This discussion is organised into three segments, the first reviewing findings from the immediate results of the study, the second discussing what can be learned about interoperability and metadata categorisation more broadly and the third looks to further work.

Immediate results

The first key finding is that metadata is a complicated medium regardless of how well structured or standardised it may be organised on the surface. As seen in the case of the DHS, both its MDCs and MDTs seem clear and offer a user a basic path through the DHS portal. Yet, conducting a deeper probe, i.e., an ontological analysis, illustrates that metadata categorisation and tagging can be subject to omissions and ambiguity. As seen in the case of the data recodification procedure, metadata cannot always capture real life events, field errors or human interactions in the form of readable indices and semantics. As outlined previously, even semantic attributes such as data variables can require further study from the user to provide further clarity, while also aiding in evading misunderstanding. Thus, a deeper, contextual description is necessary, as seen with the case of specific DHS taxonomies, outlined in 4.2.

Building upon this, in conducting this analysis, it is evident the importance that context and meaning play in facilitating true interoperability. In performing a multilevel probe of the DHS dataset portal, it is apparent that the more an analysis or study involves the interoperability of datasets in its implementation, the greater an ontological approach is required and ought to be employed. Put simply, the deeper you go, the more you'll know. Digging deeper into metadata categories at a lower level further arms the user with a better understanding of how to then implement the datasets more effectively. In the case of facilitating geographers in executing interoperability, Schuurman (2009, b) contends that metadata is an ideal mechanism for integrating non-spatial or qualitative data into GIS

systems. A further finding concerns the increasingly complexities within research methods, i.e., the increasing amount of interoperability being integrated into datasets and operational systems (e.g., the DHS's capturing of, and inclusion of locational coordinates included with non-spatial datasets, as well as spatial) means that more sophisticated metadata, i.e., layers, is necessary. Outcome from this study argues that said metadata should be more ontological in its formulation. As highlighted by Seild et al. (2018) in 2.4.1, the increasing use of spatial data by non-spatial users needs to be addressed by spatial researchers. The key issue emphasised is misinterpretation of spatial data in its description output, or classification (Seidl et at, 2018, 475, 477, 477). Thus, an ontological approach, i.e., incorporating knowledge systems within the descriptive makeup of the metadata can aid in solving this. In retrospect, the DHS have addressed this matter to a degree, for example, the creation of the spatial covariates (see 2.4.2) have allowed non-spatial users to incorporate spatial data features within their research in a manner that is validated and correct.

In addressing the second part of the research question, i.e., the facilitation of interoperability across spatial and non-spatial datasets, as a key outcome, this study recognises that there are different degrees of interoperability. As outlined in 4.3.1 (see figure 6), the level of interoperability implemented is subject to the nature of the study being conducted (e.g., a spatial analysis) and what type of datasets are intended to be employed (e.g., geographical coordinates, GIS shape files, climate indicators, etc). Therefore, metadata functionality can be applied within one dataset such as spatial data (see green intersect, figure 6, section 4.3.1), between two contrasting datasets, such as spatial and demographic data (see black intersection, figure 6, 4.3.1) or across multiple dataset types, such as spatial, demographic and health (see red intersect, figure 6, 4.3.1). Using metadata as the bridge, interoperability is enabled on various scales and across different datasets.

Furthermore, centring on the discussion outlined in 4.3.2, this study has presented three levels, or stages, afforded in applying ontological approaches to the DHS's metadata categorisation. Much like the levels of interoperability, application of ontology-based approaches are specific to the user or the intended research. In the case of this study, a level three analysis, which employed a deep contextual analysis on the DHS taxonomies and life cycle of the datasets was applicable, given that the user is inexperienced in using the DHS portal and new to its methodologies. While an experienced, well-versed user can implement the datasets and their MDCs with a level one approach, this study finds that most users, namely, geographers, could benefit in employing a level two approach of ontological practice.

This involves conducting further contextual analysis, guided by the standardisation, or DHS survey process, in understanding the dataset's origins, construction and content, and thus, better utilising them.

Subsequently, this study's findings, while fitting within the narrow scope examined and small number of datasets and MDCs analysed, acknowledges that the findings are initial and only a very small part of what would need to be a much lager and wider study, which is significantly outside the scope of this dissertation. In stating this, a primary outcome is that this investigation has illustrated that a case for an ontological approach to metadata facilitating interoperability across datasets can be made. Furthermore, using such an approach to examine the concepts of metadata and interoperability from a spatial perspective is validated, most notably within the context of evaluating a multifaceted research portal like the DHS Program. Additionally, this study has highlighted how using a bridging mechanism, i.e., an ontological approach to using metadata, can in turn, address the balance within Geography's core epistemological battle, i.e., the ideological struggle between the hard sciences, e.g., GI technologies, and the social science, tadeonal schools of thought (see 2.1).

What can be learned

This section discusses the results of the second research question, i.e., the advantages geographers could gain in using an ontological approach together with semantics when implementing metadata, in broader practice (i.e., beyond the DHS and the scope of this study). Firstly, in conducting this investigation, this study finds that the better the metadata, the more potential for interoperability. As demonstrated clearly in the case of the DHS, this study now looks to how this could impact geographers in their work. As noted in 4.3.2, Schuurman (2009b) argues metadata can be limited in its informative capacity, accounting to the low completion rates of metadata forms by users and thus, its limited utilisation. Schuurman and Leszczynski (2006, 716) support this, acknowledging in their ontology-based metadata analysis study that most spatial users do not engage with evaluating metadata, or paying heed to its descriptive outlet. Therefore, this study finds that using metadata, ideally in an ontological capacity, could tremendously benefit geographers in broadening the scope of their data studies. Furthermore, ontology, as an approach, makes for an excellent teaching tool. First time and inexperienced users of spatial data, most notably those who do not usually engage with metadata beyond a catalogue-capacity, could gain a deeper contextual understanding of the datasets to be utilised, through their collection rationale, classification method and data models (see 2.2. and 4.2), by employing an ontology-based approach.

Secondly, as seen in the case of the DHS, the utilisation of categories is effective, particularly when operating a multi-disciplinary, collaborative research platform like the DHS Program. The use of a common meta framework, i.e., the standardisation of datasets, allows researchers from different academic and scientific backgrounds (e.g., spatial, health, statistical, etc) to communicate between teams and present their work in a standardised capacity. Furthermore, the use of categories can allow researchers such as geographers to identify key research areas of interest and carrying out risk assessment. The use of multi-layered, sophisticated classification systems like the DHS's MDCs and MDTs allow researchers to swiftly assess what datasets may be sensitive, e.g., geographic coordinates, behavioural and health data, datasets relating to children, etc. The metadata systems employed by the DHS act as steppingstones, allowing researchers to search for deeper meaning behind data by using the indicators or dataset variables as their guide within the larger DHS research hub. To epitomize this, this study looks to Schuurman's (2009b) map analogy, outlined in 2.1. In seeing the world as a map, one should first look to its classification system, i.e., the map's legend. In principle, the map can only be as good in its guiding capacity, as that of its legend, the guidance to the map itself.

Further work

In conducting this study, it is evident that there is a strong convergence between the development between geography and computer science, specifically, the methods and technologies being deployed are constantly crossing over. In their study, Schuurman and Leszczynski (2006, 718-719) took the meta fields illustrated in figure 1 (see 2.2) and applied an ISO standardisation process, thus creating semantic taxonomies. Considering this, in addition to the concepts discussed in section 2.2, further work could see introducing technologies such as machine learning (ML), artificial intelligence (AI), pattern recognition and other big data analyses into the geography realm. While the DHS employs its own API (application programming interface) system, the *Indicator Data API*, this system is primarily used in as an advanced search function for data indicators and sematic metadata tags. In stating this, a future, follow-on study could investigate the possibility of creating a metadata markup language (MDML) written using XML (code classes) which could allow for a deeper meaning for the MDCs, more than just semantic code, and further made available through a standard framework. Additionally, to include a spatial dimension, the inclusion of geographer's knowledge systems and preceptive in technological developments such as these, could allow for the emergence of new sub-disciplines within the field, e.g., Smart Geography.

Further work could also look at developing a spatial markup language (SML), thus integrating geographic principles into smart technology.

It should be note that the original concept of this thesis was to look at proposing a more generic, metadata categorisation model, based on the DHS's proprietary *Survey Process* model. The research outcome intended to further analyse MDCs and thus create new, less case-specific MDCs and in turn, develop a framework for highlighting research outcomes of interest, or ROIs. These ROIs could include issues integrated within the datasets that may not be detectable from the MDCs alone, e.g., geoprivacy issues, deductive exposure, level of effective geomasking on GIS datasets, etc. Further, in analysing the DHS's use of methodology within its own research publications (see 2.4.2), a future study which aims to propose such a model, could replicate DHS methods, i.e., construct a correlation matrix ranking MDCs against potential ROIs.

5.2 Limitations

This study has considerable potential for improvement. As outlined in previous sections, this study would need to be conducted on a larger scale, examining more datasets (preferably across countries that are intercontinental, e.g., Asia, South America, etc) and more analysing metadata categories. Furthermore, a larger study would examine more of the DHS data variables and indicators, following the same framework outlined by Schuurman and Leszczynski (2006). Regarding the operational framework used in this study, Schuurman and Leszczynski's (2006) study on ontology-based metadata was an ideal model framework to follow, given that its core objective looking at how to extend current metadata schemes to include context-based information which will further enhance interoperability. Despite this, this is not to say that a paper which replicates the study design of this dissertation, i.e., examining the metadata makeup of multifaceted research portal or databased from a GIS or spatial perspective, would not have also been an ideal analysis to mirror.

Secondly, this study recognises that in using DHS resources requires a considerable amount of prior knowledge, or the time and capacity to learn how about their methodology and how to use their datasets (as discussed in 3.3.2). In addition to this, the process of applying for access to the datasets while also learning to utilise their resources (through sample datasets and surveys) is considerably time consuming. Furthermore, the size and scale

of the DHS portal and its operations must be taken into consideration. The DHS has been operating for more than forty years, across 90 different countries, deploying some 400 different surveys, thus the data collected and available to utilise is vast.

Thirdly, as outlined in the theoretical framework, metadata as a concept, is one not often examined or even considered by geographers and spatial researchers. Thus, the availability of up to date, relevant (i.e., written from a mixed-methods or qualitative approach, e.g., ontology, Critical GIS, etc) may be limited.

Chapter 6 – Conclusions

The genesis of this study came from an interest in geo-ethics and privacy relating to spatial data. The role and application of metadata not only addresses this issue but has proven to have a much wider reach. Its importance is seen both across datasets of different types (spatial, demographic, and health) and across datasets from different research studies (e.g., Tanzania and Uganda). Although literature related specifically to this topic was difficult to find and relatively sparce, the core literature did validate and endorse both the role of metadata and the approach taken in terms of the ontological method adopted and applied. A carefully applied ontological consideration of metadata clearly facilitates interoperability across spatial and non-spatial datasets.

In particular, considering the three levels of ontological analysis outlined (see 4.3.2), a well organised and carefully considered taxonomy of metadata categories was identified as being the most effective in this regard. Even establishing a basic understanding and interconnectedness between simpler metadata categories and metadata tags was shown to be useful and worthwhile.

Furthermore, the ontological approach outlined is by no means complete and would merit continued work to enhance both the breath and depth of the metadata analysis, particularly across a wider range of datasets and perhaps most challengingly yet interestingly, between research portals. Where a study such as this to be considered on a larger scale, as a future work, then this would seek to standardise the metadata categories within an ISO framework.

The identification and selection of the DHS portal as the core case study, proved to be very worthwhile, as it clearly demonstrated the central role and importance of metadata. Firstly, the metadata proved a very helpful and insightful guide to the user within the portal. Secondly, the metadata categories aided in providing context and understanding of the multifaceted and complex datasets and their collection rationale in relation to the DHS survey process. Thirdly, the DHS in its own methodology considers the metadata, from earliest stages of field work through to the final stages of dataset publication. This study considers the DHS *survey process* to be potentially pioneering and meriting adoption more broadly.

As this study has shown, spatial data is being gathered using increasingly more sophisticated collection techniques and advanced technologies, thus requiring metadata

categorisation, and tagging to become more refined in order to keep pace. This is particularly important in regard to achieving affective interoperability between increasingly sophisticated spatial and non-spatial datasets.

This study has also highlighted how metadata aids in data description, specifically in bringing a degree of context and meaning to data. As discussed in 4.1, metadata is the descriptive mechanism which enables point data on a map to be correctly and more fully interpreted as an object of meaning, such as a house. Thus, as argued in both the theoretical framework and in the study outcomes, metadata should play a more central part in spatial analyses, most notably in a qualitative, or descriptive role.

Finally, although this study presented major challenges in terms of the amount of time and commitment required to learn and understand the DHS process and methodologies, in addition to grappling with the semantic and linguistic complexities, nuances and even at times ambiguities of metadata, it has proved to be very worthwhile and a rewarding exercise. There is no doubt that metadata is going to have an important role in Geography and spatial research going forward. Systematically and methodologically embracing the full potential of metadata will help Geography to maintain its relevance in the face of an increasingly digitally enhanced world. To quote Mark Zuckerberg, Chief Executive of Facebook, “the metaverse offers an online virtual realm where people would work, play and shop” (Needleman, 2021).

References

Calvo, R.A., Deterding, S. and Ryan, R.M., 2020. Health surveillance during covid-19 pandemic.

Danko D., 2008. Metadata and Interoperability, Geospatial. In: Shekhar S., Xiong H. (eds) Encyclopedia of GIS. Springer, Boston, MA. https://doi-org.ludwig.lub.lu.se/10.1007/978-0-387-35973-1_780.

Diamond, C.C., Mostashari, F. and Shirky, C., 2009. Collecting and sharing data for population health: a new paradigm. *Health affairs*, *28*(2), pp.454-466.

Dorélien, A., Balk, D. and Todd, M., 2013. What is urban? Comparing a satellite view with the demographic and health surveys. *Population and Development Review*, *39*(3), pp.413-439.

Dobson, J.E. and Fisher, P.F., 2003. Geoslavery. IEEE Technology and Society Magazine, 22(1), pp.47-52.

Elwood, S., Schuurman, N. and Wilson, M., 2011. Critical Gis. *The SAGE Handbook of GIS and Society. London, Sage Publications Ltd*, pp.87-106.

Elwood, S. and Leszczynski, A., 2011. Privacy, reconsidered: New representations, data practices, and the geoweb. *Geoforum*, *42*(1), pp.6-15.

Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., Parker, M., Bonsall, D. and Fraser, C., 2020. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science*, *368*(6491).

Goodchild, M. F., 1992. Geographical information science. *International journal of geographical information systems*, *6*(1), 31-45.

GPSA (Global Partnership for Social Accountability). 2020. “Making Public Contracts Work for People: Experiences from Uganda.” GPSA. Available at: https://www.thegpsa.org/stories/making-public -contracts-work-people-experiences-uganda. (Accessed: 2021/07/25).

Grace, K., Nagle, N.N., Burgert-Brucker, C.R., Rutzick, S., Van Riper, D.C., Dontamsetti, T. and Croft, T., 2019. Integrating environmental context into DHS analysis while protecting participant confidentiality: A new remote sensing method. *Population and development review*, *45*(1), p.197.

ICF. 2018. Demographic and Health Surveys Standard Recode Manual for DHS7. The Demographic and Health Surveys Program. Rockville, Maryland, U.S.A.: ICF. Available at: https://www.dhsprogram.com/pubs/pdf/DHSG4/Recode7_DHS_10Sep2018_DHSG4.pdf (Accessed: 2021/07/25).

ICF International. 2013. *Incorporating Geographic Information into Demographic and Health Surveys: A Field Guide to GPS Data Collection.* Calverton, Maryland, USA: ICF International. Available at: https://dhsprogram.com/pubs/pdf/DHSM9/DHS_GPS_Manual_English_A4_24May2013_DHSM9.pdf (Accessed: 2021/07/25).

Kostkova, P., 2018. Disease surveillance data sharing for public health: the next ethical frontiers. Life sciences, society and policy, 14(1), p.16.

Kwan, M., 2007. Affecting Geospatial Technologies: Toward a Feminist Politics of Emotion. The Professional Geographer no 59 (1) pp. 27-34.

Longley, P.A., Goodchild, M., Maguire, D. J., & Rhind, D. W., 2015. *Geographic Information Systems and Science*. John Wiley & Sons.

Ministry of Health, Community Development, Gender, Elderly and Children - MoHCDGEC/Tanzania Mainland, Ministry of Health - MoH/Zanzibar, National Bureau of Statistics - NBS/Tanzania, Office of Chief Government Statistician - OCGS/Zanzibar, and ICF. 2016. Tanzania Demographic and Health Survey and Malaria Indicator Survey 2015-2016. Dar es Salaam, Tanzania: MoHCDGEC, MoH, NBS, OCGS, and ICF. Available at http://dhsprogram.com/pubs/pdf/FR321/FR321.pdf. (Accessed: 2021/04/01).

Ministry of Health (MOH) [Uganda] and ORC Macro. 2006. *Uganda HIV/AIDS Sero-behavioural Survey 2004-2005*. Calverton, Maryland, USA: Ministry of Health and ORC Macro. Available at https://dhsprogram.com/pubs/pdf/AIS2/AIS2.pdf (Accessed 2021/07/25).

Needleman, S.E., 2021. 'Mark Zuckerberg Sketches Out Facebook's Metaverse Vision', The Wall Street Journal, 28th October. Available at: https://www.wsj.com/articles/mark-zuckerberg-to-sketch-out-facebooks-metaverse-vision-11635413402 (Accessed: 2021/10/28).

Nouwt, S., 2008. Reasonable expectations of geo-privacy. *SCRIPTed*, *5*, p.375.

O'Sullivan, D., 2006. Geographical information science: critical GIS. *Progress in Human Geography*, 30(6), 783-791.

Pavlovskaya, M., 2018. Critical GIS as a tool for social transformation. *The Canadian Geographer/Le Géographe Canadien*, *62*(1), pp.40-54.

Pavlovskaya, M. and Martin, K.S., 2007. Feminism and geographic information systems: From a missing object to a mapping subject. *Geography Compass*, *1*(3), pp.583-606.

Pavlovskaya, M., 2006. Theorizing with GIS: a tool for critical geographies?. *Environment and Planning A*, *38*(11), pp.2003-2020.

Percivall G., 2008. OGC's Open Standards for Geospatial Interoperability. In: Shekhar S., Xiong H. (eds) Encyclopedia of GIS. Springer, Boston, MA. https://doi-org.ludwig.lub.lu.se/10.1007/978-0-387-35973-1_904

Rambaldi, G., Chambers, R., McCall, M. and Fox, J., 2006. Practical ethics for PGIS practitioners, facilitators, technology intermediaries and researchers. *Participatory learning and action*, *54*(1), pp.106-113.

Sinton, D.S., 2017. How Do We Integrate Ethics into Geospatial Education?.

Schuurman, N., 2009**a**. *Critical GIS*. Encyclopedia of Human Geography vol. 2 pp. 363-368.

Schuurman, N., 2009**b.** Metadata as a site for imbuing GIS with qualitative information. *Qualitative GIS: a mixed methods approach*, pp.40-56.

Schuurman, N. and Leszczynski, A., 2006. Ontology-based metadata. *Transactions in GIS*, *10*(5), pp.709-726.

Schuurman, N., 2005. Object Definition in GIS. *Re-presenting* GIS, p.27. – 2.1.2

Schuurman, N., 2000. Trouble in the heartland: GIS and its critics in the 1990s. *Progress in human geography*, *24*(4), pp.569-590.

Smith, B., and Mark, D.M., 1998. Ontology and geographic kinds.

Seidl, D.E., Jankowski, P. and Nara, A., 2019. An empirical test of household identification risk in geomasked maps. *Cartography and Geographic Information Science*, *46*(6), pp.475-488.

Seidl, D.E., Paulus, G., Jankowski, P. and Regenfelder, M., 2015. Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography*, *63*, pp.253-263.

Standard DHS 2015-2016 [Dataset] TZMR7B.SAV. Ministry of Health, Community Development, Gender, Elderly and Children - MoHCDGEC/Tanzania Mainland, Ministry of Health - MoH/Zanzibar, National Bureau of Statistics - NBS/Tanzania, Office of Chief Government Statistician - OCGS/Zanzibar, and ICF. 2016. Tanzania Demographic and Health Survey and Malaria Indicator Survey 2015-2016. Dar es Salaam, Tanzania: MoHCDGEC, MoH, NBS, OCGS, and ICF. [Producers]. ICF [Distributor], 2016.

Swanlund, D. and Schuurman, N., 2019. Resisting geosurveillance: A survey of tactics and strategies for spatial privacy. *Progress in Human Geography*, *43*(4), pp.596-610.

Thatcher, J.E., Bergmann, L. and O'Sullivan, D., 2018. Speculative and constructively critical GIS.

The DHS Program, 2017. Uganda Bureau of Statistics (UBOS) and ICF. 2017. 2016 *Uganda Demographic and Health Survey Key Findings.* Kampala, Uganda, and Rockville, Maryland, USA. UBOS and ICF. Available at: https://dhsprogram.com/pubs/pdf/SR245/SR245.pdf (Accessed 2021/07/25).

The DHS Program, 2016. Ministry of Health, Community Development, Gender, Elderly and Children (MoHCDGEC), [Tanzania Mainland, Ministry of Health (MoH) [Zanzibar], National Bureau of Statistics (NBS), Office of the Chief Government Statistician (OCGS) and ICF. 2016. 2015-16 TDHS-MIS Key Findings. Rockville, Maryland, USA: MoHCDGEC, MoH, NBS, OCGS, and ICF. Available at https://dhsprogram.com/pubs/pdf/SR233/SR233.pdf (Accessed: 2021/04/01).

The DHS, (n.d.a). Team and Partners. The DHS [online], c2013. Available at: https://www.dhsprogram.com/Who-We-Are/About-Us.cfm (Accessed: 2021/04/01).

The DHS, (n.d.b). Survey Process. The DHS [online], no date. Available at: https://www.dhsprogram.com/Methodology/Survey-Process.cfm (Accessed: 2021/04/01).

The DHS, (n.d.c) *Survey-Process-Design*. [online] Available at: https://dhsprogram.com/Methodology/Survey-Process.cfm (Accessed: 2021/04/01).

Trinadh D., Assaf, S., Yourkavitch, J. and Mayala, B. 2018. *A Primer on The Demographic and Health Surveys Program Spatial Covariate Data and Their Applications.* DHS Spatial Analysis Reports No. 16. Rockville, Maryland, USA: ICF.

Uganda Bureau of Statistics - UBOS and ICF. 2018. *Uganda Demographic and Health Survey 2016.* Kampala, Uganda and Rockville, Maryland, USA: UBOS and ICF. Available at: https://dhsprogram.com/pubs/pdf/FR333/FR333.pdf (Accessed: 2021/07/25).

Uganda Bureau of Statistics - UBOS and ICF. 2018. *Uganda Demographic and Health Survey 2016.* Kampala, Uganda and Rockville, Maryland, USA: UBOS and ICF. Available at: https://dhsprogram.com/pubs/pdf/FR333/FR333.pdf (Accessed 2021/07/25).

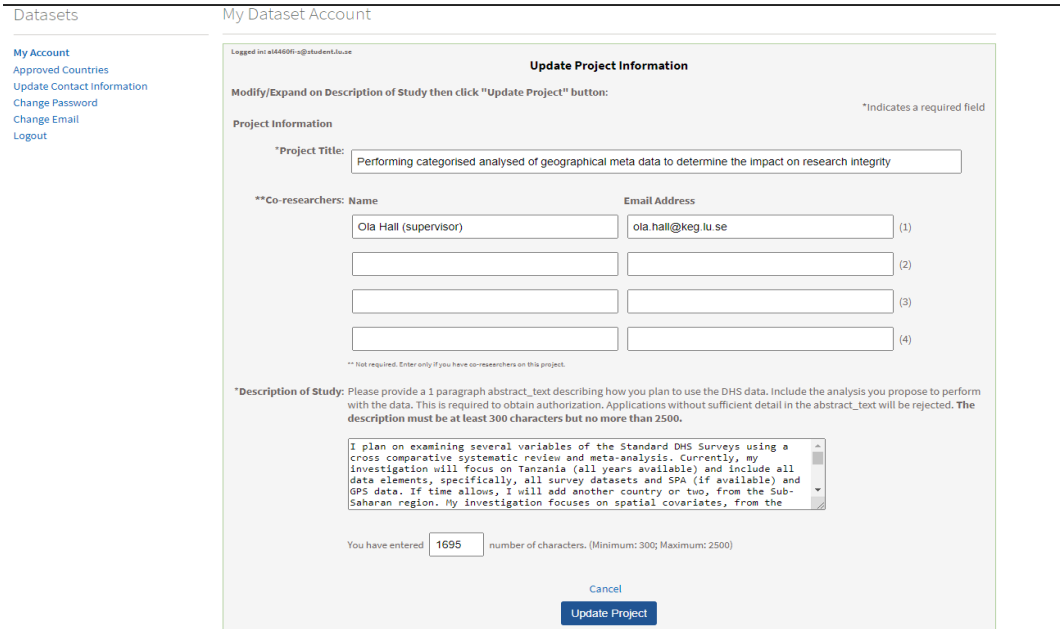
Weiser, P. and Scheider, S., 2014, November. A civilized cyberspace for geoprivacy. In *Proceedings of the 1st ACM SIGSPATIAL international workshop on privacy in geographic information collection and analysis* (pp. 1-8).

World Bank. 2021. World Development Report 2021: Data for Better Lives. Washington, DC: World Bank. doi:10.1596/978-1-4648-1600-0. License: Creative Commons Attribution CC BY 3.0 IGO.

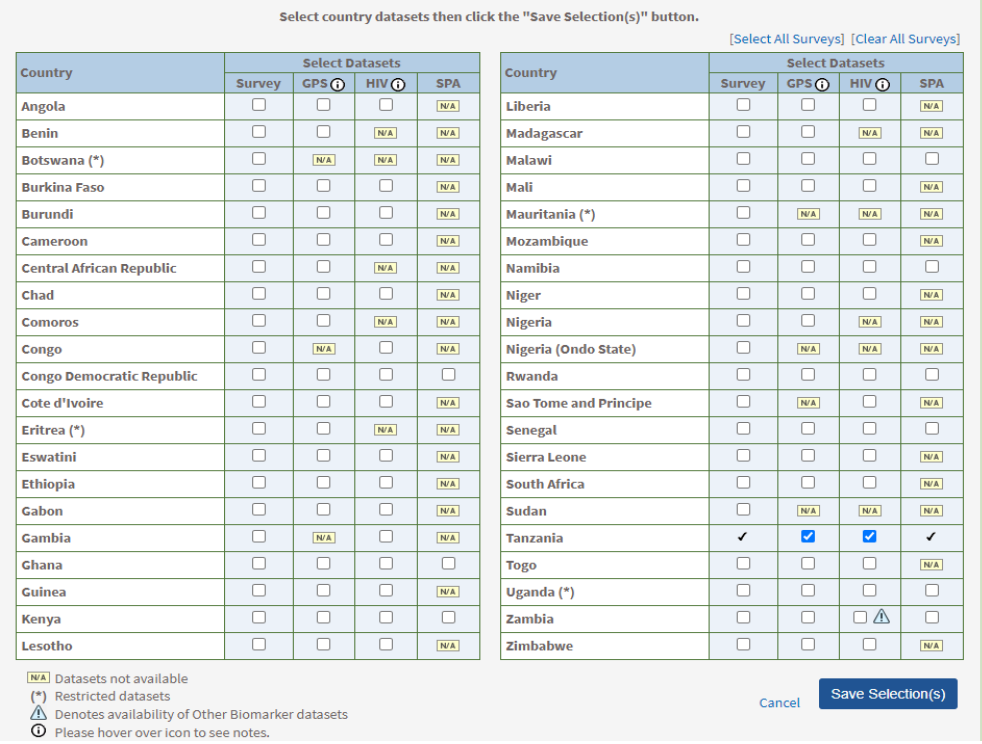
Appendix 1

Screenshots for application DHS process.

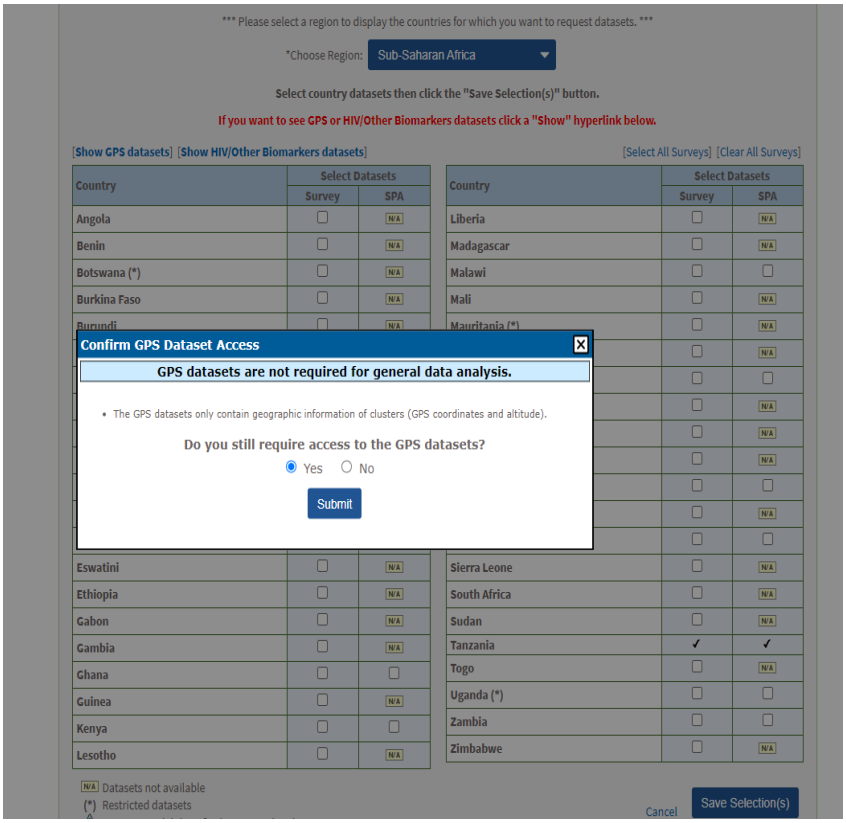
Tanzania:



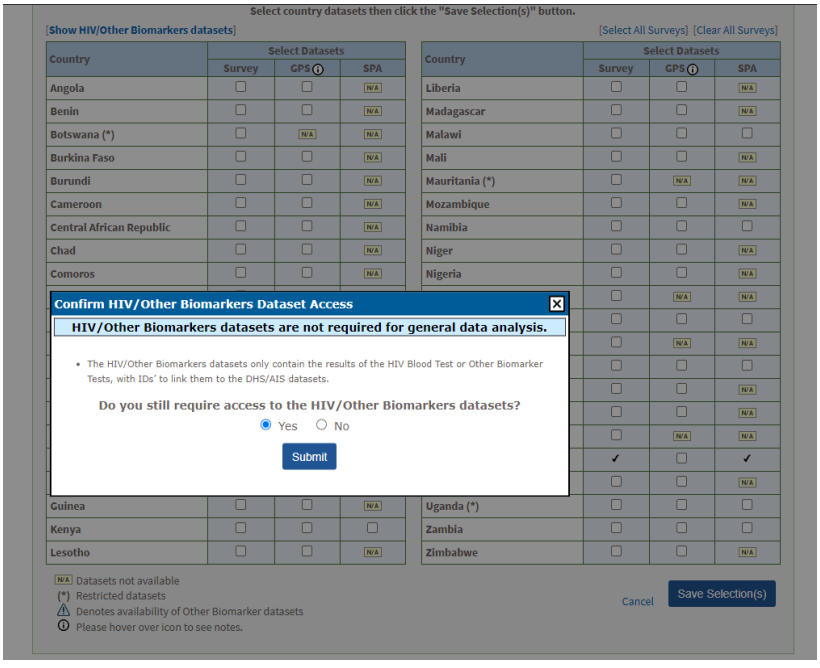
Screenshot 1: Research proposal



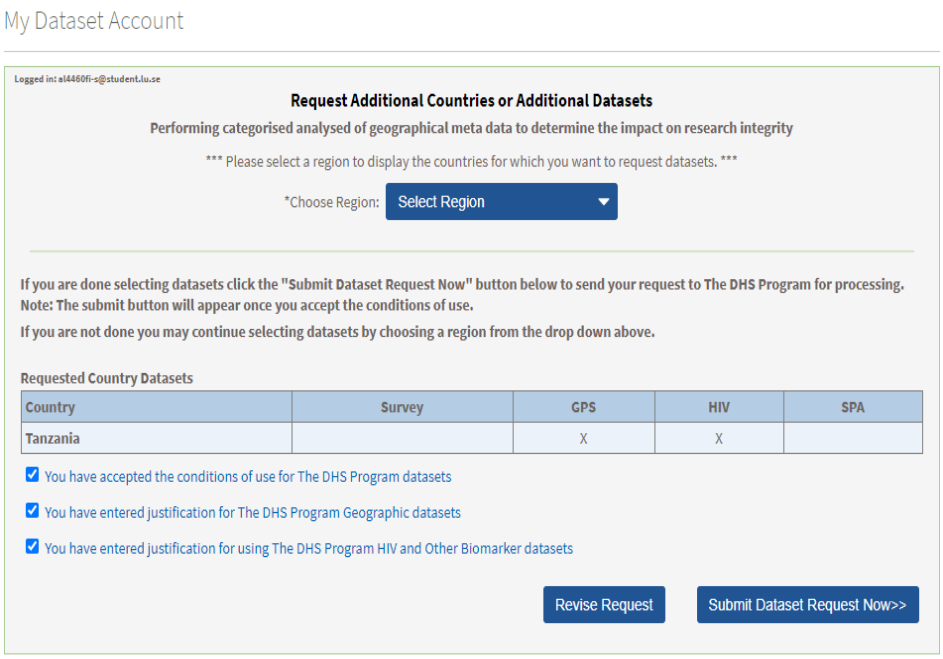
Screenshot 2: Selection of country and its available datasets



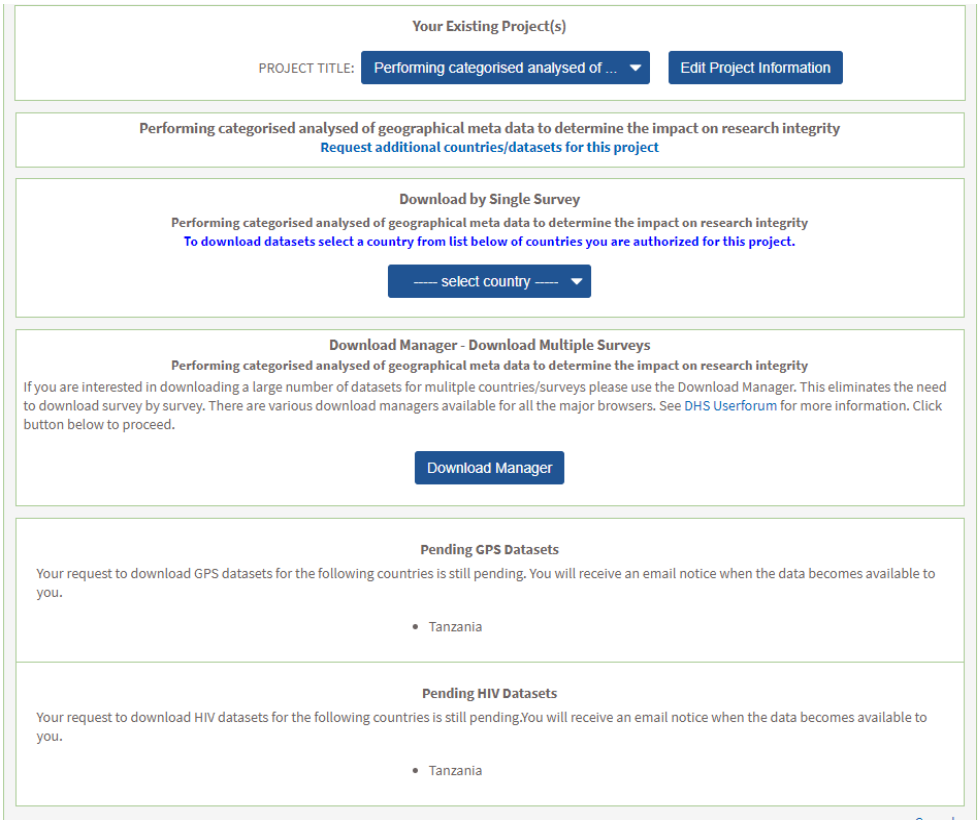
Screenshot 3: Specific selection of GPS datasets



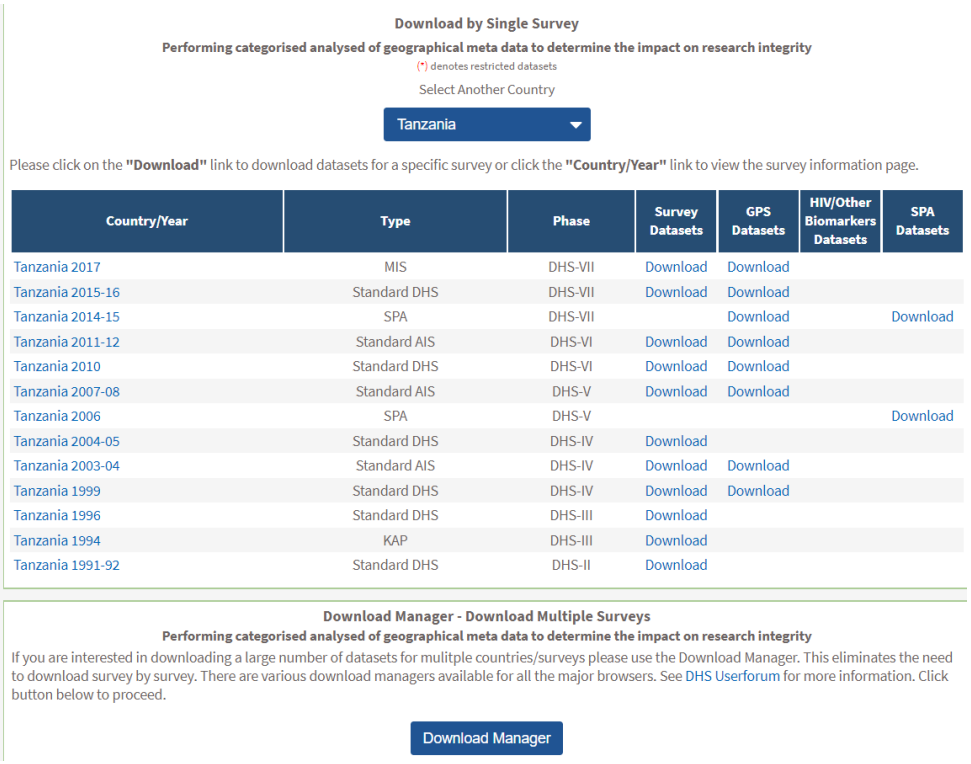
Screenshot 4: Specific selection of HIV/Other datasets



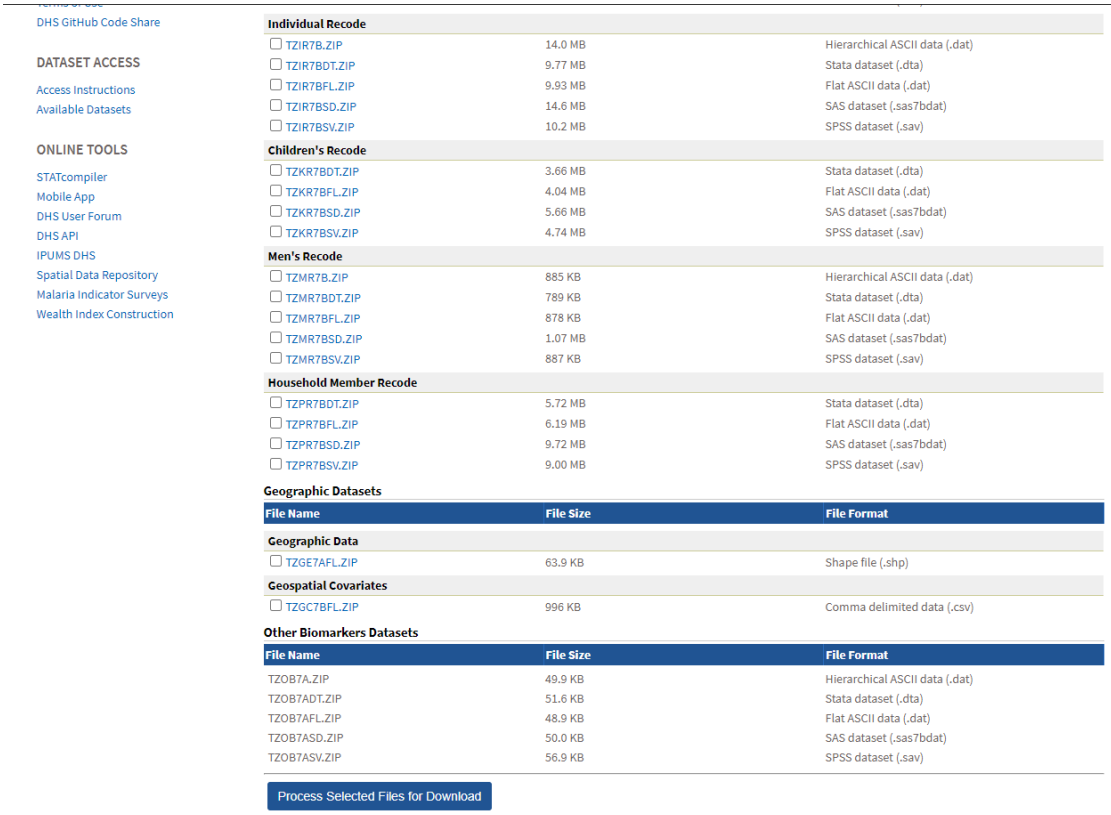
Screenshot 5: Requesting additional (non-standard) datasets – includes accepting the T&Cs in addition to filling out extra justifications for use of other datasets



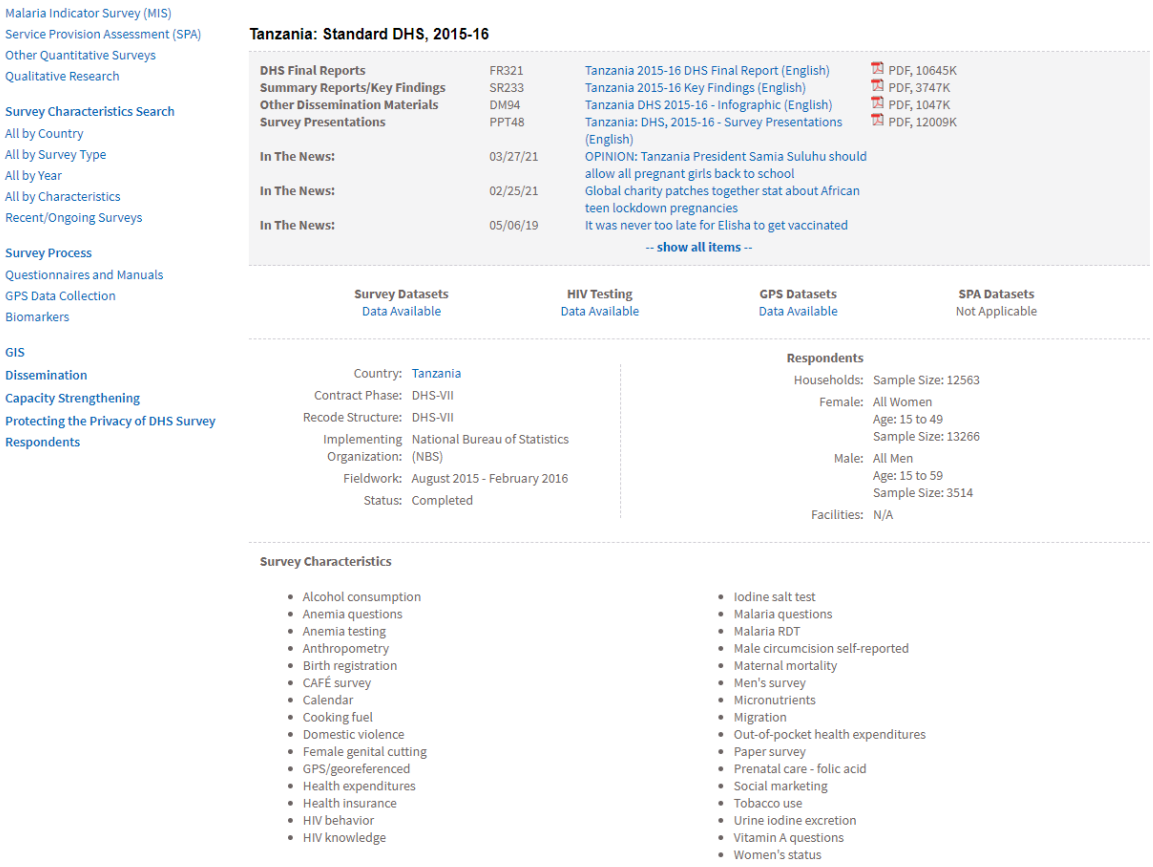
Screenshot 6: Dataset(s) approval pending



Screenshot 7: Datasets for Tanzania (approved)

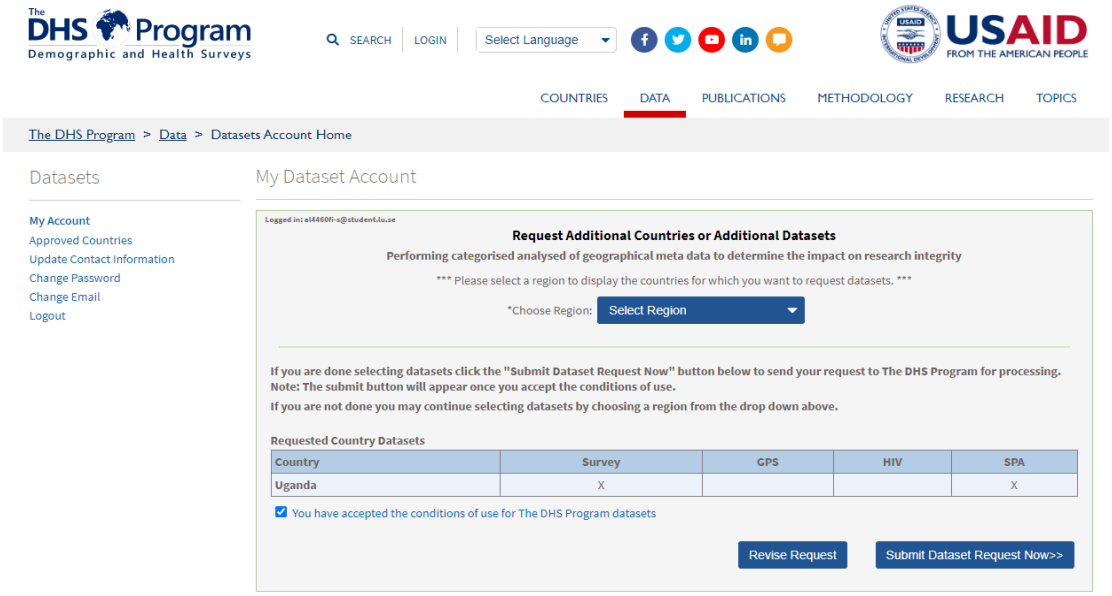


Screenshot 8: Tanzania Standard DHS, 2015-2016 dataset files and format for download. Note the inclusion of GPS datasets and biomarkers, complentmentary to survey datasets.

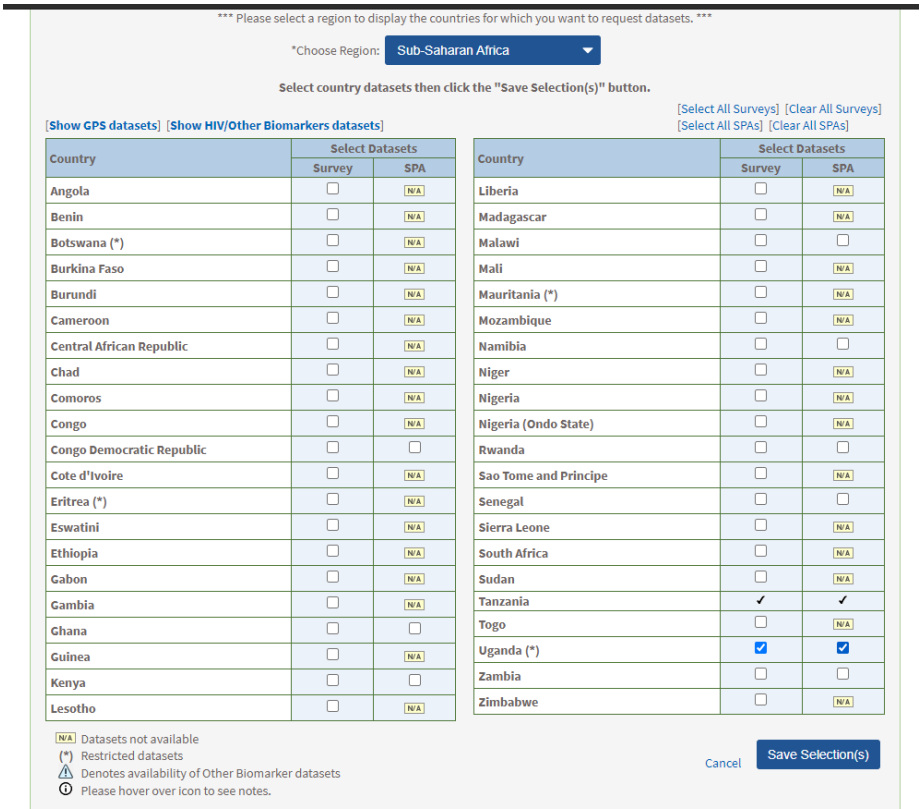


Screenshot 9: Summary of the Tanzania Standard DHS, 2015-2016 dataset. This includes the dataset types, survey charactersitics (what was asked) and reports …

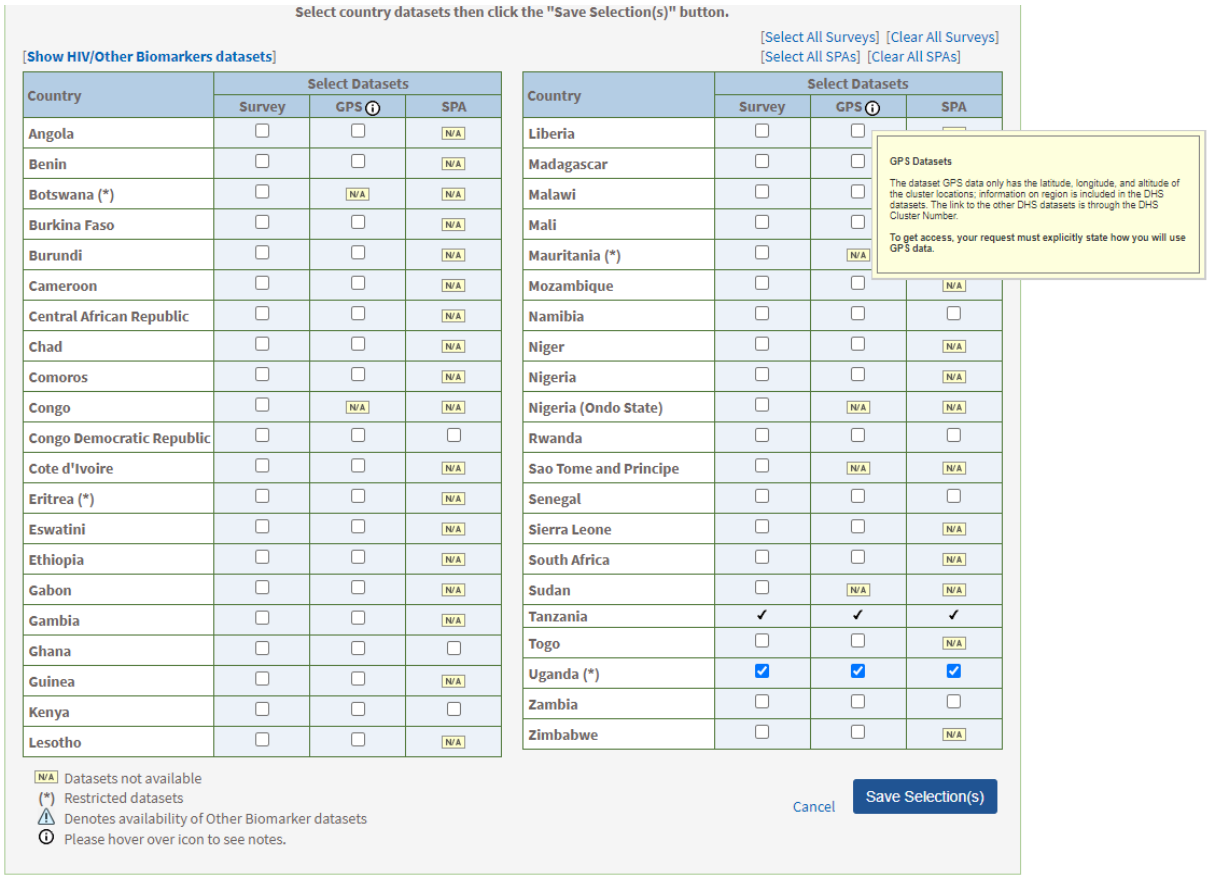
Uganda:



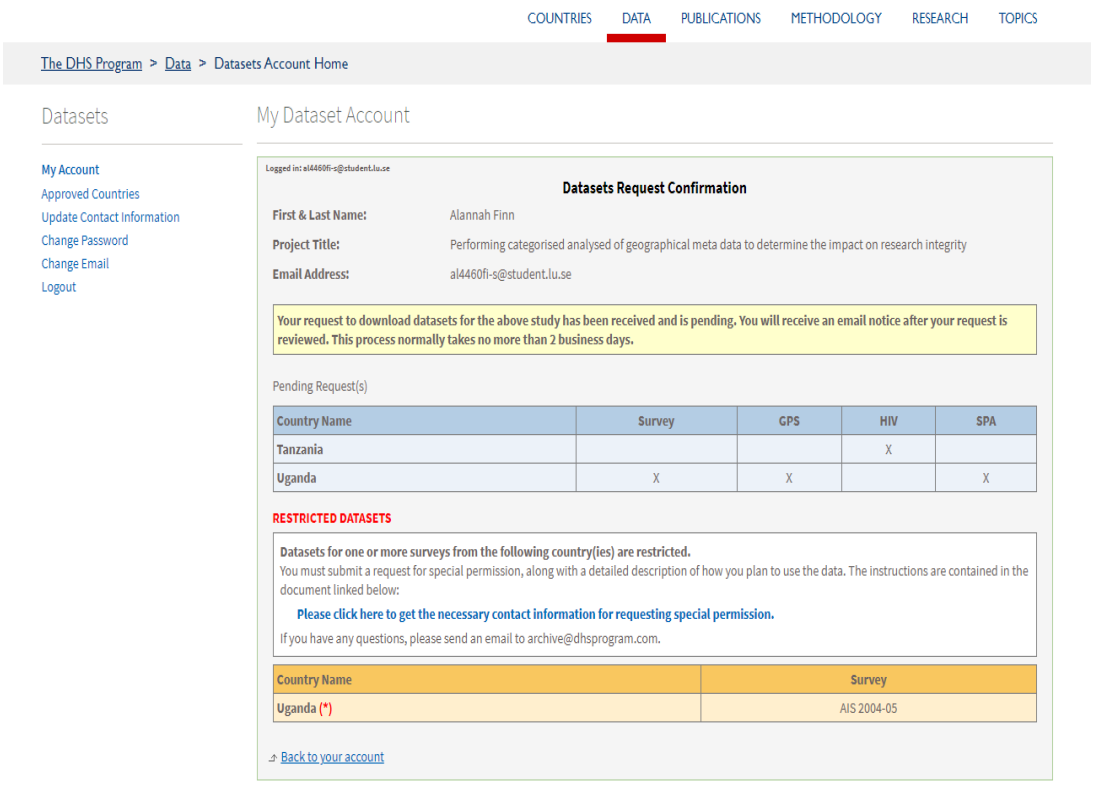
Screenshot 10: Requesting an additional country: Uganda standard datasets selection



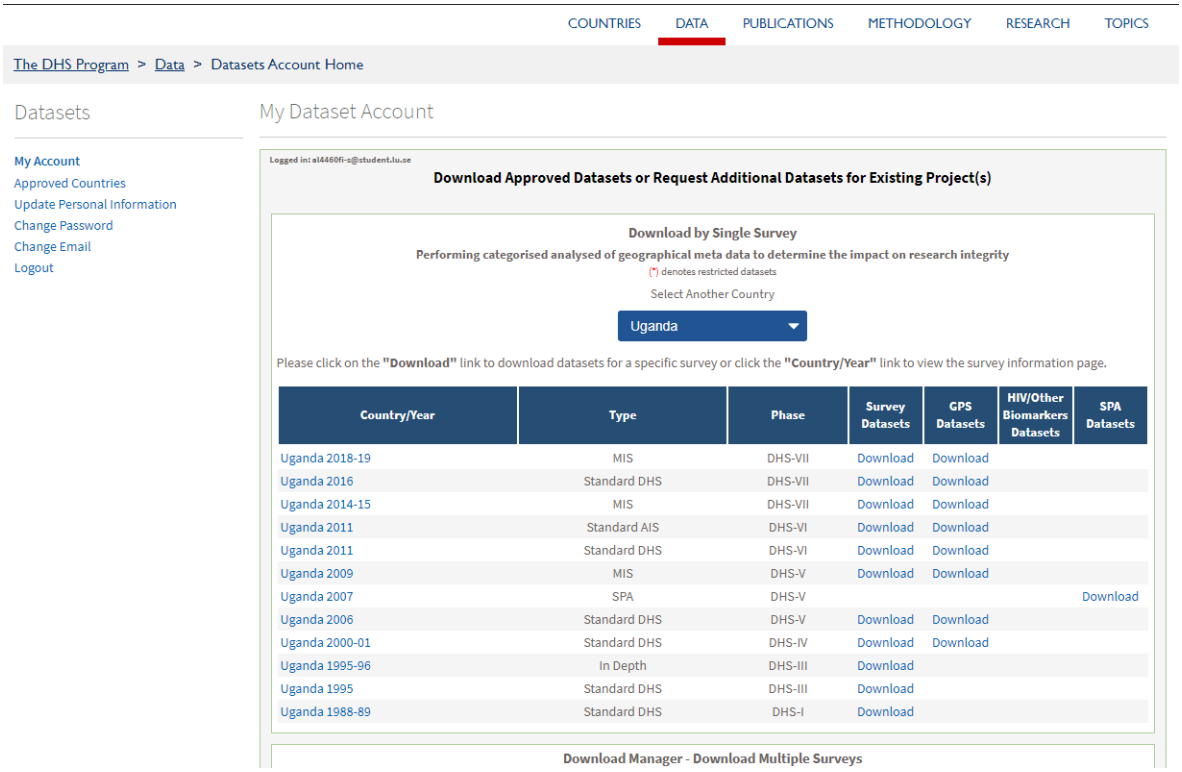
Screenshot 11: DHS: Data has restricted access for Uganda (see * marker)



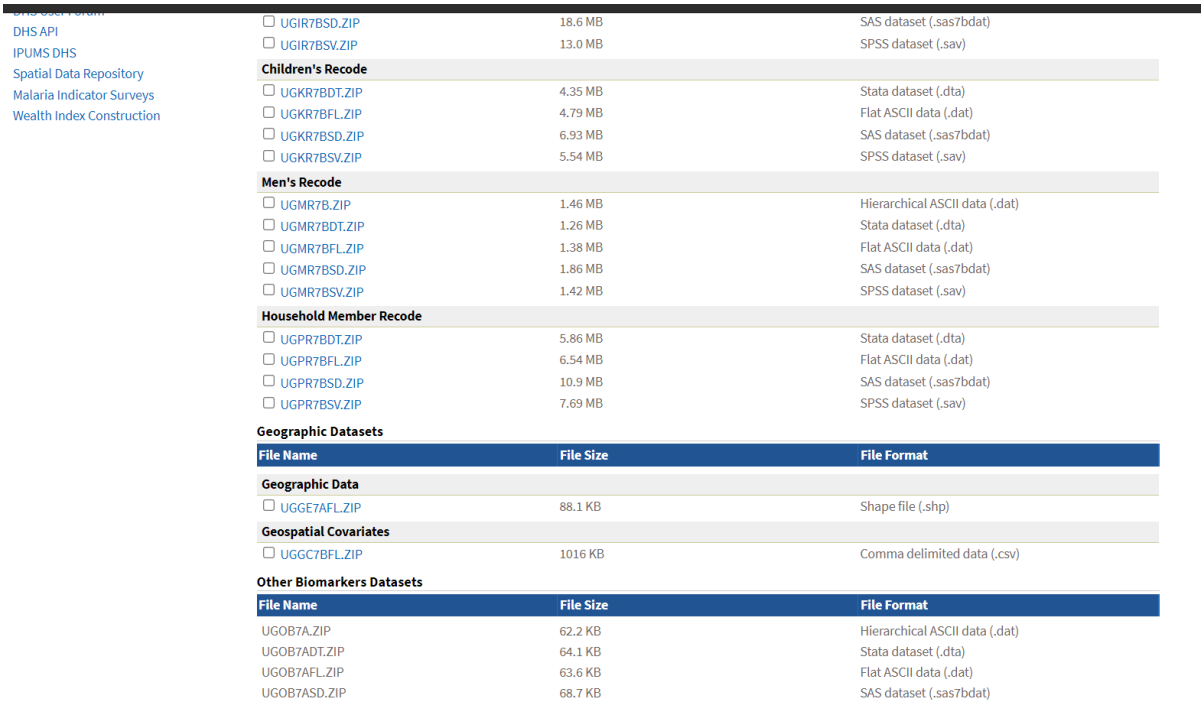
Screenshot 12: Selection request for additional (non-standard) GPS and SPA data for Uganda



Screenshot 13: Request pending… Includes notes on restricted access to certain datasets (requires additional request and clearance)



Screenshot 14: Approved datasets for Uganda



Screenshot 15: Uganda Standard DHS, 2015-2016 dataset files and format for download (Includes GPS datasets and biomarkers).

Appendix 2

Links to reports/dataset citations.

Tanzania:

Report 1: Tanzania DHS, 2015-16 - Final Report (630 pages)

https://dhsprogram.com/pubs/pdf/FR321/FR321.pdf

Report 2: Tanzania DHS, 2015-2016 – Key Findings report [summary of Final Report] (28 pages)

https://dhsprogram.com/pubs/pdf/SR233/SR233.pdf

Dataset citation: Standard DHS 2015-2016 [Dataset] TZMR7B.SAV. Ministry of Health, Community Development, Gender, Elderly and Children - MoHCDGEC/Tanzania Mainland, Ministry of Health - MoH/Zanzibar, National Bureau of Statistics - NBS/Tanzania, Office of Chief Government Statistician - OCGS/Zanzibar, and ICF. 2016. Tanzania Demographic and Health Survey and Malaria Indicator Survey 2015-2016. Dar es Salaam, Tanzania: MoHCDGEC, MoH, NBS, OCGS, and ICF. [Producers]. ICF [Distributor], 2016.

Uganda:

Report 3: Uganda DHS, 2016 - Final Report (625 pages)

https://dhsprogram.com/pubs/pdf/FR333/FR333.pdf

Report 4: Uganda DHS, 2016 – Key Findings report [summary of Final Report] (Uganda DHS, 2016) (20 pages)

https://dhsprogram.com/pubs/pdf/SR245/SR245.pdf

Report 5: Uganda HIV/AIDS Sero-Behavioural Survey 2004-05 [Findings from restricted datasets]

https://dhsprogram.com/pubs/pdf/AIS2/AIS2.pdf