Common Language for Accessibility, Interoperability, and Reusability in Historical Demography

Rick J. Mourits[1][0000−0002−2267−1679], Tim Riswick[2][0000−0003−1401−6284], and Rombert Stapel[3][0000−0001−6394−260X]

[1] International Institute of Social History, Amsterdam, Netherlands
rick.mourits@iisg.nl
[2] Radboud University, Nijmegen, Netherlands
tim.riswick@ru.nl
[3] International Institute of Social History, Amsterdam, Netherlands
rombert.stapel@iisg.nl

Abstract. One of the biggest challenges in the transition to open science is making data interoperable. Ideally, existing schemas and vocabularies are (re-)used to describe data, but these are generally problematic for historical data, as they exclude historical concepts and are insensitive to temporal variations in meaning. Therefore, the subdiscipline of historical demography has designed its own schemas and vocabularies to standardize historical data, as researchers require them to make and study large-scale reconstructions of populations and life courses. We introduce a web environment called CLAIR-HD that helps researchers to find vocabularies to standardize historical demographic data, and determine lacunae in the standardization of data within the field of historical demography.

Keywords: Interoperability · Vocabularies · Historical Demography

1 CLAIR-HD

One of the biggest hurdles in data interoperability is communication. Without coordination, database managers tend to come up with different descriptions for the same information, which hurts data interoperability. To tackle this problem, vocabularies and schemas are used to standardize how data in datasets is being described. Sometimes these standardization efforts are very straightforward and apply to very broad contexts, whereas others are of general use to specific communities. For historical data, however, most of these standardization efforts are problematic as they were made to describe contemporary concepts and underappreciate how information and meaning can change over time. For example, places and their names change over time, the associations between occupation and social standing shift when labour markets change, and causes of death are coded differently between social and temporal contexts [17–19, 23, 24, 34, 35, 47]. Existing vocabularies standardize these historical concepts at the cost of losing or misinterpreting context-specific information.

In this paper, we use the subdiscipline of historical demography as a case study to see how specialized vocabularies are made and adopted. Historical demographers from a wide array of countries have built databases to reconstruct the lives of people in Europe, North America, and East Asia [22]. The vocabularies of these databases were designed to "stay true to the source", so that datasets have sophisticated designs to model local peculiarities and changes in meaning over time. These local efforts have made it possible to standardize defunct phenomena, historical distinctions, and general changes over time, though only within the geographic scope of their projects. Each of these standardization schemes is worth its weight in gold, as it unlocks a wealth of historical data and contains years of insight in the historical sources and context. Yet, standardized communication is necessary to make larger-scale comparisons possible. Currently, the field is in a paradoxical situation where most scholars agree that historical comparisons increase understanding of the "historical context", but are at the same time afraid to throw out the baby with the bathing water and hesitant to apply schemes not designed for a specific social, spatial, and temporal context [24, 33, 54].

To have fruitful discussions on vocabularies, historical demographers need to have an overview of existing standardization efforts. Yet, this requires too much effort for most scholars, as they need to know the field well, have expertise in presenting data, and invest time in ontology design. To reduce the time and knowledge required to partake in this discussion, we gathered the vocabularies that data centers, projects, and research collaborations developed to make data interoperable and variables comparable. To show how these vocabularies are related, we mapped the relationships between them and created an overview of vocabulary conversion tools. These results are published on the CLAIR-HD webpage, so that researchers and database managers can easily find and reuse existing vocabularies. Gathering and sharing these vocabularies helps historical demographers to learn from each other's insights, prevents the re-invention of vocabularies, and ensures that data is interoperable. But most importantly, it serves as a case study on what is required for a move towards open data within history, and perhaps even the humanities and social sciences as a whole, as common vocabularies allow for general-purpose software, make replication studies easier, and are the steppingstone to Linked Open Data. We intend for CLAIR-HD to become an inspiration to other disciplines that face similar challenges.

2 Methods

Information on standardization efforts within historical demography was gathered in multiple rounds. Our initial goal was to get a broad outline of the existing vocabularies. Therefore, we contacted the bigger data centers in Asia, Europe, and North America. These data centers were a logical place to start, as they have the most developed infrastructure and are important regional hubs in historical demography. Although all historical data centers responded promptly, it turned out that few disclosed their own standardization lists and often reused existing vocabularies to standardize their data. Nevertheless, there were notable exceptions and paradoxically, the data centers often worked together to develop shared code books, also known as schemas. Hence, we decided to also gather information on the schemas that they developed. This gave us a feeling for how data standardization efforts in the field were designed and why there has been less focus on developing shared vocabularies than initially anticipated.

We also talked to individual scholars at the major historical demography conferences in Europe and North America: European Social Science History Conference (ESSHC), European Society for Historical Demography (ESHD), and the Social Science History Association conference (SSHA). Here we learned which vocabularies historical demographers use to code their data and how they work together to interpret and codify historical data. In general, researchers seem to prefer using multiple vocabularies in their analyses to test whether differences in interpretation and standardization practices can lead to different statistical associations, either to look for overlap as a robustness check or to tease out differences as an enquiry into underlying mechanisms. Therefore, we decided to also collect information on crosswalks and conversion tools, as they hugely increase data interoperability and give researchers an important tool in their toolbox.

3 Existing standardization efforts

The outcomes of the enquiry are interpreted in this section of the paper and presented on the CLAIR-HD web page and section 4. Our initial goal was to get a broad outline of the existing vocabularies and show the overlap between them. Once the data came in, it became clear that vocabularies are generally products of scholarly collaborations. Whether these vocabularies are used is dependent on the quality of the product itself, other scholars familiarity with it, and willingness of data providers to implement it. To understand the institutional context within which vocabularies are being provided, we first describe the seven schemas that were developed within and for historical demography. Second, we give an overview of the existing vocabularies and determine lacunae in the standardization of data within the field of historical demography. Finally, we close by discussing crosswalks and conversion tools.

3.1 Schemas

Our inventory of the field showed that each data center uses its own schema. However, there are also seven schemas that provide standardized ways to deliver datasets, 1. the Intermedidate Data Structure (IDS) [1, 2], 2. IPUMS-USA [40], 3. IPUMS-International [25], 4. LINKS-gen [29], 5. MOSAIC [46], 6. North Atlantic Population Project (NAPP) [36–38], and 7. Persons in Context (PiCo) [5].

Theme	**Variable**	**IPUMS**	**MOSAIC**	**NAPP**
Geography	Country code	CNTRY	country	CNTRY
	Place	-	place	-
	Region	region	place	-
	Urban-rural	URBAN	urban	URBAN
Household	Group quarter status	GQ	gq	GQ
	Household size	PERSONS	hhsize	NUMBERHH
	Household weight	WTHH	hhwt	HHWT
Identifier	Enumeration	SAMPLE	id_enum	SAMPLE
	Household	SERIAL	id_hhold	SERIAL
	Person	PERNUM	id_pers	PERNUM
Individual	Age	AGE	age	AGE
	Literacy	LIT	lit	LIT
	Marital status	MARST	marst	MARST
	Occupational title	-	occupan	OCCSTR
	OCCHISCO	OCCHISCO	occhisco	OCCHISCO
	Present at enum.	RESIDENT	presence	RESIDENT
	Rel. household head	-	relate	-
	Religion	RELIG	relig	RELIGION
	Sex	SEX	sex	SEX
	Weight	WTPER	perwt	PERWT
Person name	First name	-	fname	NAMEFRST
	Last name	-	lname	NAMELAST
Quality	Age	-	qage	QAGEGB
	Household	-	qhhold	-
	Rel. household head	-	qrelate	QRELGB
	Marital status	-	qmarst	QMARSTGB
	SEX	-	qsex	QSEXGB
Source	Enumeration type	-	enumtype	-
	Enumeration year	YEAR	year	YEAR

Table 1: Census schemas (IPUMS-international, MOSAIC, NAPP) [46]

Although all these schemas are meant to standardize historical data, their intended scope differs. IPUMS-USA, IPUMS-International, MOSAIC, and NAPP were developed to standardize census data. The driving force behind processing census data is IPUMS at the University of Minnesota. In 1991, they started providing "common-format extracts" with standardized codes and constructed variables for the 1960, 1970, and 1980 US censuses and now maintain standards to exchange census data within the USA (IPUMS-USA) and internationally (IPUMS-International) [39]. In 1999, IPUMS joined up with researchers from Canada, Great Britain, Iceland, Norway, and Scotland with whom they already had established strong ties. As they realised that the original source material was highly compatible and cultural constructs for the measured concepts are similar, they decided to create a machine-readable, census-based database of, as they put it, the North Atlantic world at the end of the 19th century [36, 38].

A similar international census comparison project took place in the early 2010s at the Max Planck Institute for Demographic Research (MPIDR) in Rostock, Germany. The MPIDR schema, called MOSAIC, standardized census data from the 1700s until 1950 for 18 regions in Europe [46]. These schemas to describe census data are very similar as shown in Table 1. Yet, all four schemas are still in use, as they cater to a very specific public and have not been adopted by other historical demographic research projects or databases, as they have been developed for projects with strong institutional boundaries.

Whereas IPUMS-USA, IPUMS-International, MOSAIC, and NAPP focus on standardizing categories between series of cross-sectional census data, other schemas set up standards for sharing person reconstructions. IDS, LINKS-gen, and PiCo are efforts to standardize historical data from other types of historical sources, such as the civil registry, militia registers, parish registers, population registers, slave registers, or tax registers. Of these schemas, LINKS-gen is by far the most limited in its scope and standardizes historical data into a pedigree format - with each row representing a person that is linked to his family by links to one's father, mother, and spouses - and a standardized occupational table that are both ready for statistical analysis [29]. IDS has been around since 2009 and makes different types of data sources available for extraction by explicitly stating for which point or period in time historical information is valid [1, 2]. PiCo goes a step further and is developed by the Center for Family History in the Netherlands as a means to store information on persons registrations as well as concomitant records and person reconstructions [5]. Table 2 shows how different these schemas are from each other, as well as from the IPUMS, MOSAIC, and NAPP census schemas. As LINKS-gen and PiCo are relatively new, it is still uncertain to what extent they will be implemented by other historical demographic research projects and databases, which will ultimately determine their longevity.

All schemas within historical demography deal with many similar concepts, but have very limited interoperability. This resemblance is understandable as most schemas are designed for tabular datasets with historical person data, hence one would expect that most schemas contain a standardized variable name and categories for common concepts that describe historical persons and the relations between them. Yet, historical data centers have been deeply rooted in national research traditions and are more focused on disclosing historical sources, matching data, and reconstitution families than on exchanging data [8, 9, 21, 22, 42]. As a result, variable names and categories are generally standardized within one institutional context rather than by using a shared vocabulary within the field of historical demography. This is even true for concepts that can be standardized very easily, such as birth or marriages dates, which have no common name or set date format to order day, month, and year. This general lack of a common language for historical demography means that schemas only function within their institutional context, are hard to find for people not actively looking for them, and require a plethora of conversion tools to move between them.

Variable	**IDS**	**LINKS-gen**	**PiCo**
Baptism	BAPTISM_DATE	-	-
Birth	BIRTH_DATE	B_date	schema:birthDate
Date	TIMESTAMP	Date	-
Death	DEATH_DATE	D_date	schema:deathDate
Divorce	DIVORCE_DATE	-	-
First obs	START_OBSERVATION	-	-
Funeral	FUNERAL_DATE	-	-
Last obs	START_OBSERVATION	LastEntryDate	-
Marriage	MARRIAGE_DATE	M_date_n	-
Mar. banns	MARRIAGE_PROCLAMATION_DATE	-	-
Stillbirth	STILLBIRTH_DATE	-	-
Baptism	BAPTISM_LOCATION	-	-
Birth	BIRTH_LOCATION	B_location	schema:birthPlace
Death	DEATH_LOCATION	D_location	schema:deathPlace
Divorce	DIVORCE_LOCATION	-	-
Funeral	FUNERAL_DATE	-	-
Marriage	MARRIAGE_LOCATION	M_location_n	-
Mar. banns	MARRIAGE_PROCLAMATION_LOCATION	-	-
Place	-	Location	schema:address
Stillbirth	STILLBIRTH_LOCATION	-	-
Relation to other person	RELATION	-	-
ID father	ID_I_2	Id_father	schema:parent
ID household	ID_C	-	-
ID mother	ID_I_2	Id_mother	schema:spouse
ID partner	ID_I_2	Id_partner_n	schema:spouse
ID person	ID_I_1	Id_person	pico:personObservation
Age	AGE_YEARS AGE_MONTHS AGE_WEEKS AGE_DAYS	Age	pico:hasAge
Age at death	-	D_age	-
Age last obs.	-	LastEntryAge	-
Alive / dead	ALIVE	-	pico:deceased
HISCO	OCCUPATION_HISCO	HISCO	-
Mar. status	CIVIL_STATUS	-	-
Marriages	-	Marriages_N	-
Nationality	NATIONALITY	-	-
Occ. title	OCCUPATION	Occupation	schema:hasOccupation
Religion	RELIGION	-	pico:hasReligion
Role	-	-	pico:hasRole
Sex	SEX	Sex	schema:gender
Twin	MULTIPLE_BIRTH	Twin	-
Archive	-	-	schema:holdingArchive
Image	-	-	schema:url
Source	-	-	prov:hadPrimarySource

Table 2: Date, geography, household, identifiers, individual, and source information in the IDS, LINKS-gen, and PiCo schemas

Taking a closer look at PiCo shows that there is a way to solve the problems with findability, accessability, and interoperability by reusing existing schemas from domains outside historical demography. In line with insights from earlier projects [24], the designers behind PiCo describe concepts in their schema with vocabularies from existing schemas to make their data more FAIR, as shown in Table 3. They do so by describing their data in different layers. A first layer is very general and uses concepts that are used around the world, such as gender or parenthood. These concepts are derived from schema.org [15]. Domain-specific concepts, such as biological events or data provenance, are borrowed from domain-specific schemas like BIO and PROV-O [6, 58]. Finally, concepts that are specific to historical demography or the civil registry are defined in specialized vocabularies, in this case their own PiCo vocabulary. As a result, PiCo contains the best practices from existing schemas and reuses them to be more easily findable, accessible on the internet, and interoperable with other databases.

However, simply reusing vocabularies is not enough. The authors of PiCo run into the same problem as the designers of other schemas when they start to describe concepts unique to the field of historical demography. Following their own design logic, they ought to implement vocabularies from accepted schemas within historical demography. But a widely accepted schema with concomitant vocabularies is simply not around. PiCo tried to solve this issue by defining its own definitions, putting it at risk of also becoming bound to a specific institutional context. Nevertheless, PiCo presents part of the solution to making historical data FAIR and less constraint by institutional boundaries. The other half of the solution is provided by collaborating researchers whose grassroot initiatives have resulted in specialized vocabularies on concepts such as cause of death categories and social status.

Schema level	Schema name	Concept name
General	XSD [11]	date, int, string
	Schema.org [15]	spouse, parent, gender, familyName givenName, ArchiveComponent, dateCreated, locationCreated
Domain-specific	BIO [6]	Marriage, date, partner
	PROV-O [58]	hadPrimarySource, wasDerivedfrom
Specialized	PiCo [5]	PersonObservation, hasRole, huwelijkspartij, huwelijksakte, hasAge

Table 3: Concentric description of a marriage certificate using PiCo

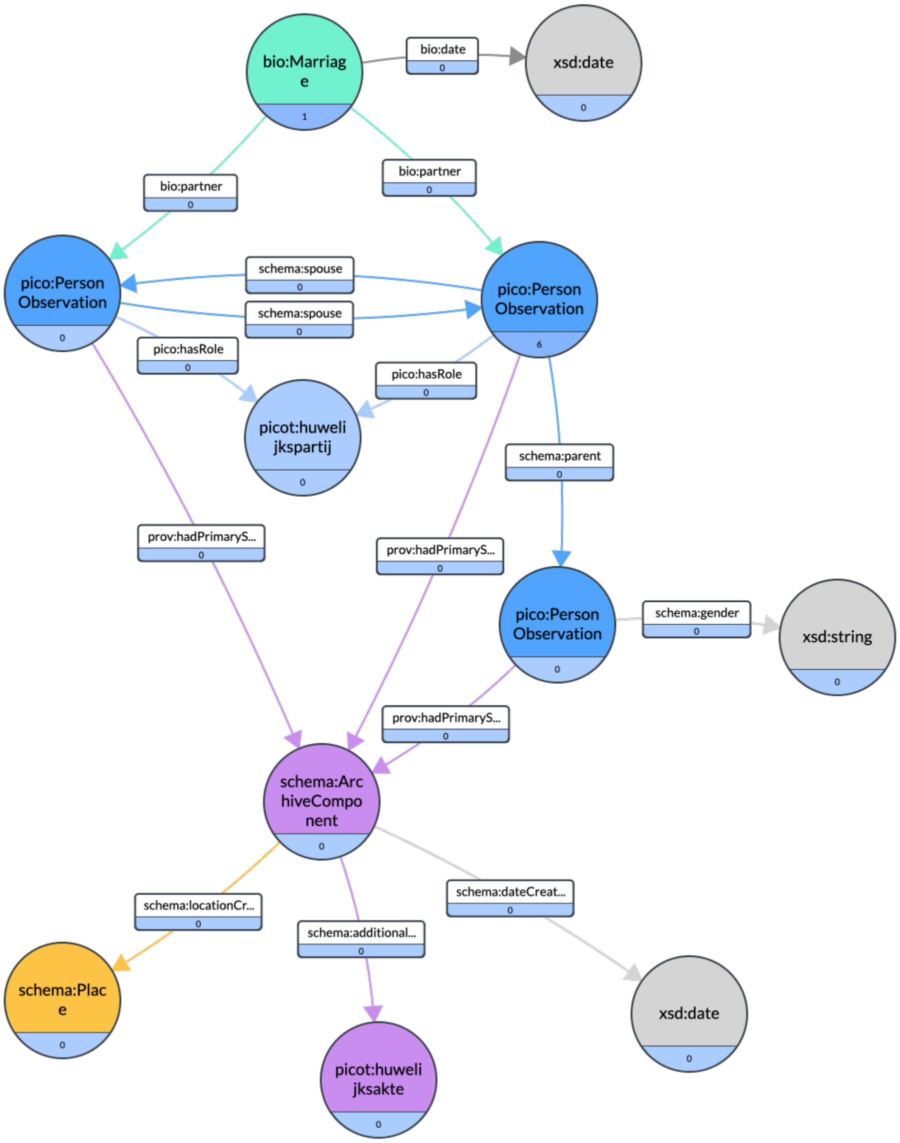


Figure 1: Marriage certificate according to the PiCo model [48]

3.2 Vocabularies

Currently, eleven concepts in historical demography are described with publicly available, standardized vocabularies, 1. cause of death titles, 2. cause of death groupings, 3. occupational titles, 4. occupational groupings, 5. social status codes, 6. place names, 7. grouped geographically defined (administrative) areas, 8. titles of religious denominations, 9. groupings of religious denominations, 10. (household) relations, and 11. data quality flags, as shown in Table 4.

Topic	**Standardized titles**	**Groupings**	**Status codes**
Cause of death	[17, 45]	[17, 45]	-
Occupations	[4, 10, 20, 26, 32, 45]	[37, 49, 55]	[7, 19, 31, 41, 50, 53]
Place names	[12–14, 16]	[43, 44, 52]	-
Religious denominations	[28, 51]	[3]	-
(Household) relations	[1, 25, 27, 40, 46]	-	
Data quality flags	[1, 25, 27, 40, 46]	-	-

Table 4: Available vocabularies within historical demography

Occupational entries are by far the most standardized. Researchers have been using occupational status schemes for several decades, which has resulted in clear pipelines for processing occupational information. Coding occupational information generally consists of three steps. First, entry errors, abbreviations, and spelling variations are removed to standardize occupational titles [4, 10, 20, 26, 32]. Second, these occupational titles are grouped into occupational groups using intermediate coding schemes. In Europe this is generally HISCO, a system developed by two historical sociologists [49], whereas the standards in the USA are OCC1950 and OCC1990, two systems developed by the United States Bureau of the Census [55]. In a third step, these occupational groups are assigned occupational status codes, such as Duncan's socioeconomic index [7], HISCLASS [50], HISCAM [19], Nam-Powers-Boyd occupational scores [30, 31], Siegel prestige score [41], SOCPO [53], and other social status measures. Table 5 provides an example from the SwedPop database.

The coding system behind occupational titles shows that three steps are required for coding historical concepts. 1. standardization of titles, 2. grouping into codes, and 3. assigning status codes. The arduous nature of these steps is shown by the SHiP project, which aims to standardize cause of death titles for multiple countries in Europe and codify them [17]. The project received funding to build a historical causes of death network out of existing collaborations. Over the past six years, researchers from around Europe have been working together to standardize causes of death titles and categorize them. Their goal is to develop a vocabulary that can "deal well with large numbers of historical disease descriptions, from different linguistic areas in Europe, while at the same time it is able to connect to current day disease patterns". The fruits of their labour are expected to be presented later in 2024.

The amount of effort that was invested in creating a historical International Classification of Diseases (ICDh) also turned out to be its biggest strength. Over the past few conferences anticipation has slowly been building. A year before its launch, ICD10h has already been accepted as the de facto standard for standardizing and coding historical cause of death titles, as a wide range of researchers contributed. Individual efforts to standardize historical information, such as the Linked International Classification for Religions (LICR) [3], have been far less successful. SHiP shows that specialized vocabularies only succeed if multiple scholars come together as a network, commit time to exchange expertise, and create excitement for their vocabularies.

OCCUPATION STANDARD	**HISCO**	**S**	**R**	**P**	**HISCAM**
FARTYGSARBETARE	98100	-9	-9	-9	65
FD TREDJE KLASS FARTYGSARBETARE	98100	-9	21	-9	65
BÅTFÖRMAN	98120	-9	-9	-9	60
BÅTFÖRMANÄNKA	98120	-9	11	-9	60
BÅTKARLFÖRMAN	98120	-9	-9	-9	60
FD BÅTFÖRMAN	98120	-9	21	-9	60
FISKEBÅTSFÖRMAN	98120	-9	-9	-9	60
VATTENBÅTFÖRMAN	98120	-9	-9	-9	60
SJÖFÖRMAN	98130	31	-9	-9	53
ANDRA KLASS SJÖMAN	98135	-9	-9	-9	53
ANDRA KLASS SJÖMAN VID FLOTTAN	98135	-9	-9	-9	53
ANDRA KLASS SJÖMANHUSTRU	98135	-9	11	-9	53
...					
BESÄTTNINGSLÄRLING	98140	33	-9	-9	53
BESÄTTNINGSPOJKE	98140	-9	-9	-9	53
BÅTDRÄNG	98140	-9	-9	-9	53
DÄCKMATROS	98140	-9	-9	-9	53
...					
BOGSERBÅTSBESÄTTNINGSKARL	98190	-9	-9	-9	46
BÅTBITRÄDE	98190	-9	-9	-9	46
BÅTFÖRARE	98190	-9	-9	-9	46
BÅTFÖRAREARBETARE	98190	-9	-9	-9	46

Table 5: Excerpt from the SwedPop standardization, grouping, and status assignment of occupational codes using HISCAM [45]
S, R, and P refer to the HISCO status, relation, and product code [19, 45, 49].

It is important to realize that work on specialized vocabularies can focus on international as well as longitudinal comparisons. However, local comparisons over time generally require a higher level of detail. Work on standardization of place names and geographically defined (administrative) areas shows that discussions do not necessarily have to take place in international contexts. While modern place name vocabularies are widely available [12, 13, 56], for historical place names around the world we are less spoilt for choice, although significant effort is being made [14]. These place name vocabularies, or gazetteers, are rarely the best solution to refer to historical spaces, as administrative areas tends to change over time. Therefore, historical demographers have matched place names to geographically and temporally defined geometries, which are generally made available as shape files. In turn, these geometries can be used to calculate spatial statistics [43, 44, 52].

It is beyond the scope of this article to deeply delve into the relationship between place names and associated geographic units. However, researchers with strong ties to spatial demography are doing their best to model the underlying complexities within countries and assign shape files to administrative areas. For example, the Amsterdamse code (AMCO) was developed for the Netherlands to solve issues with applying modern coding systems for municipalities to historical settings [52]. This system works well for the 19th and 20th centuries when administrative areas were more or less fixed. However, a more flexible semantic model was necessary to refer to the more fluid premodern administrative areas within the Low Countries [43, 44]. Such regional or national geographical efforts are generally more useful than one-size-fits-all solutions that span the globe, as they allow users to define how spatial units should be grouped. Yet, such flexibility should not come at the cost of intelligibility. Therefore, it may be worthwhile to explore the use of discrete global grids to create a system for making intermediary layers of historical administrative areas, so that national insights can be translated to the international community. For example, to allow comparisons between Dutch municipalities [52] and Swedish parishes [45].

For other historical concepts, large-scale discussions on how to standardize titles and categorize have not started yet. As a result, data on religion, (household) relations, and data quality flags is much less standardized. Shared vocabularies are either not available or not accepted within the field, so database managers and researchers use their own categories. What is currently needed is to have the local experts join forces in a network to combine these insights in specialized vocabularies. The existence of local coding system means that future efforts to develop specialized vocabularies for religion, (household) relations, and data quality do not have to start from scratch. However, the SHiP project shows that simply having expertise is not enough, as transforming local coding systems into a shared vocabulary takes time and collaborative effort.

3.3 Conversion tools

There are currently no tools available to move between schemas. However, a sizable number of conversion tools exists to move between vocabularies. All but one conversion tools are available for occupation-related vocabularies, making it by far the most vibrant ecosystem. The only other conversion tool groups causes of death titles into the ICD10h.

Table 6 lists the conversions tools that are currently publicly available. There are tools for three different processes: to group titles into categories, to move between two systems of categorization, and to assign status codes to categories. Different people provide these conversion tools. Tools to categorize titles are generally provided by data centers who produce them as auxiliary data [4, 10, 20, 26, 32, 45]. This allows them to assign status codes using schemas that were developed by scholars [49, 50, 53], while crosswalks between occupational groupings were made by individual scholars in order to compare European, American, or project- specific social status definitions [27, 59].

In order to make conversions possible, conversions should be as straightforward as possible. Nevertheless, conversions from grouped titles to status codes can require additional information on the context. For example, HISCLASS requires information on whether labourers live in an urban or rural environment [50], useful categorizations of cause of death differ by the age of the deceased [17], and conversions from place names to geographically defined (administrative) units require an observation year. The easiest way for conversion tools to assign such context-specific status codes is a rule-based software scripts. However, the problem with scripts is that they need regular maintenance, are not intelligible to all scholars, and can easily become pretty complex. Therefore, the better practice is to make conversion tables, as they are are low-maintenance, easy to understand, and limit the number of ways in which data can be split up.

The availability of conversion tools is indicative of an environment in which researchers and database managers share best practices. Sharing conversion tools and their underlying methodology prevents redundant work and ascertains data quality. Nevertheless, there is still a sizable number of conversion tools that are not publicly available. Multiple institutes have crosswalks that are often shared upon request, but hidden from view to possibly interested researchers and database mangers. Sharing these conversion tools is low-hanging fruit in improving the interoperability of historical demographic data.

Process	**From**	**To**	**Author**
Categorize titles	Cause of death title	ICD10h	[17, 45]
Categorize titles	Occupational title	HISCO	[4, 10, 20, 26, 32, 45]
Convert groupings	HISCO	OCC1950	[27]
Convert groupings	OCC1950	HISCO	[27]
Convert groupings	HISCO	OCCHISCO	[59]
Assign status	HISCO	HISCAM	[20]
Assign status	HISCO	HISCLASS	[20]
Assign status	HISCO	SOCPO	[20]

Table 6: Publicly available conversion tools in historical demography

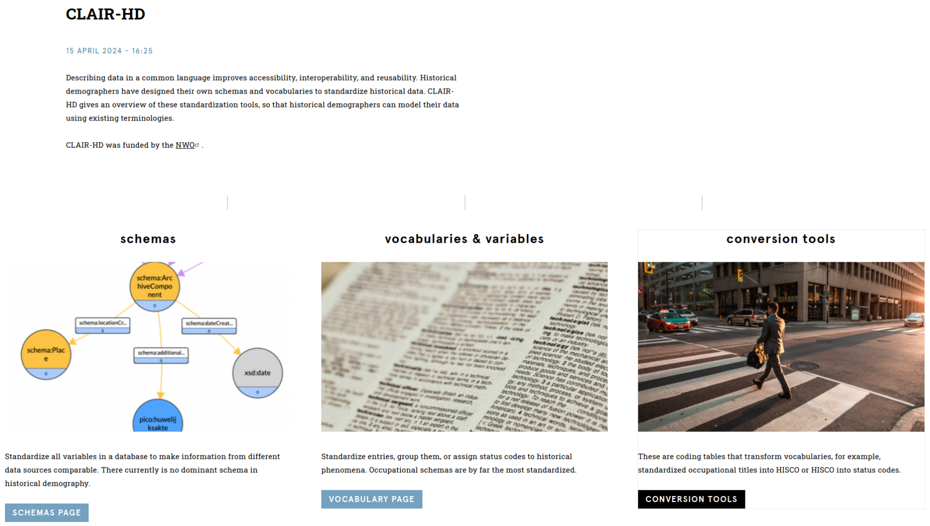
4 CLAIR-HD website

To help the field in adopting open science practices, we created a website to inform researchers about existing schemas, vocabularies, and vocabularies. The web pages are hosted and maintained by the International Institute of Social History (IISH) in Amsterdam to ascertain that these pages remain accessible in the foreseeable future. Moreover, all provided material is also stored in the IISH data repository to improve its reusability. The website can be found at: https://iisg.amsterdam/en/blog/clair-hd

Figure 2 shows the CLAIR-HD landing page which contains a brief explanation on the used terminology and links to pages with more information. This page links to a schema page, a vocabulary page, and a conversion tool page. The schema page summarizes the design principles of the different schemas (see Figure 3). IPUMS, MOSAIC, and NAPP are compared in one table to see how these schemas name variables relating to the household, identifier, individual, person name, provenance, quality indicators, and the source. A second table provides a similar comparison for IDS, LINKS-gen, and PiCo. Combined, these two tables give a quick overview of the schemas' different design principles.

The vocabulary page explains the three stadiums of standardization and presents standardization efforts per concept (see Figure 4). Per variable, efforts to standardize variables are introduced as running text and are accompanied by paragraphs on available titles, groupings, and social status codes.

The conversion tool page lists the existing tools and contains links to GitHub, institutional webpages, and repositories where these tools are hosted (see Figure 5).



CLAIR-HD

15 APRIL 2024 - 16:25

Describing data in a common language improves accessibility, interoperability, and reusability. Historical demographers have designed their own schemas and vocabularies to standardize historical data. CLAIR-HD gives an overview of these standardization tools, so that historical demographers can model their data using existing terminologies.

CLAIR-HD was funded by the NWO .

schemas

Standardize all variables in a database to make information from different data sources comparable. There currently is no dominant schema in historical demography.

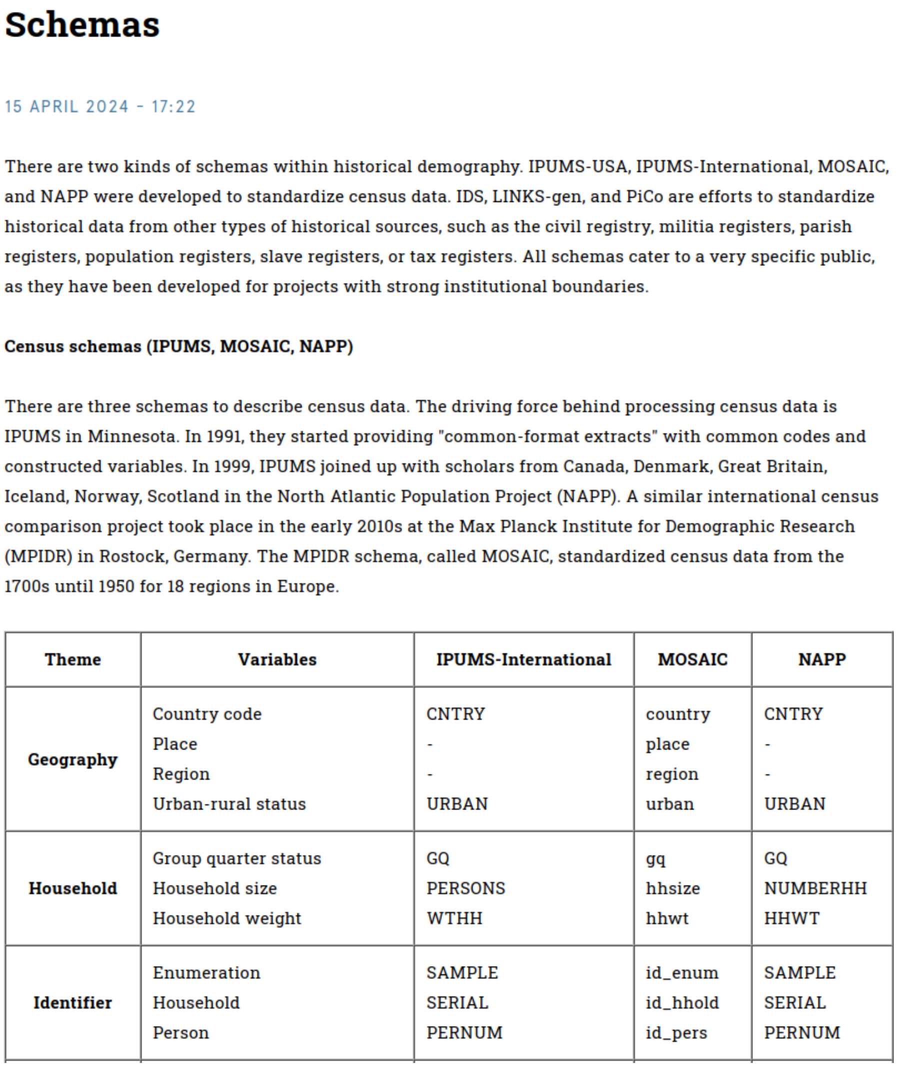
vocabularies & variables

Standardize entries, group them, or assign status codes to historical phenomena. Occupational schemas are by far the most standardized.

conversion tools

These are coding tables that transform vocabularies, for example, standardized occupational titles into HISCO or HISCO into status codes.

Figure 2: CLAIR-HD landing page



Schemas

15 APRIL 2024 - 17:22

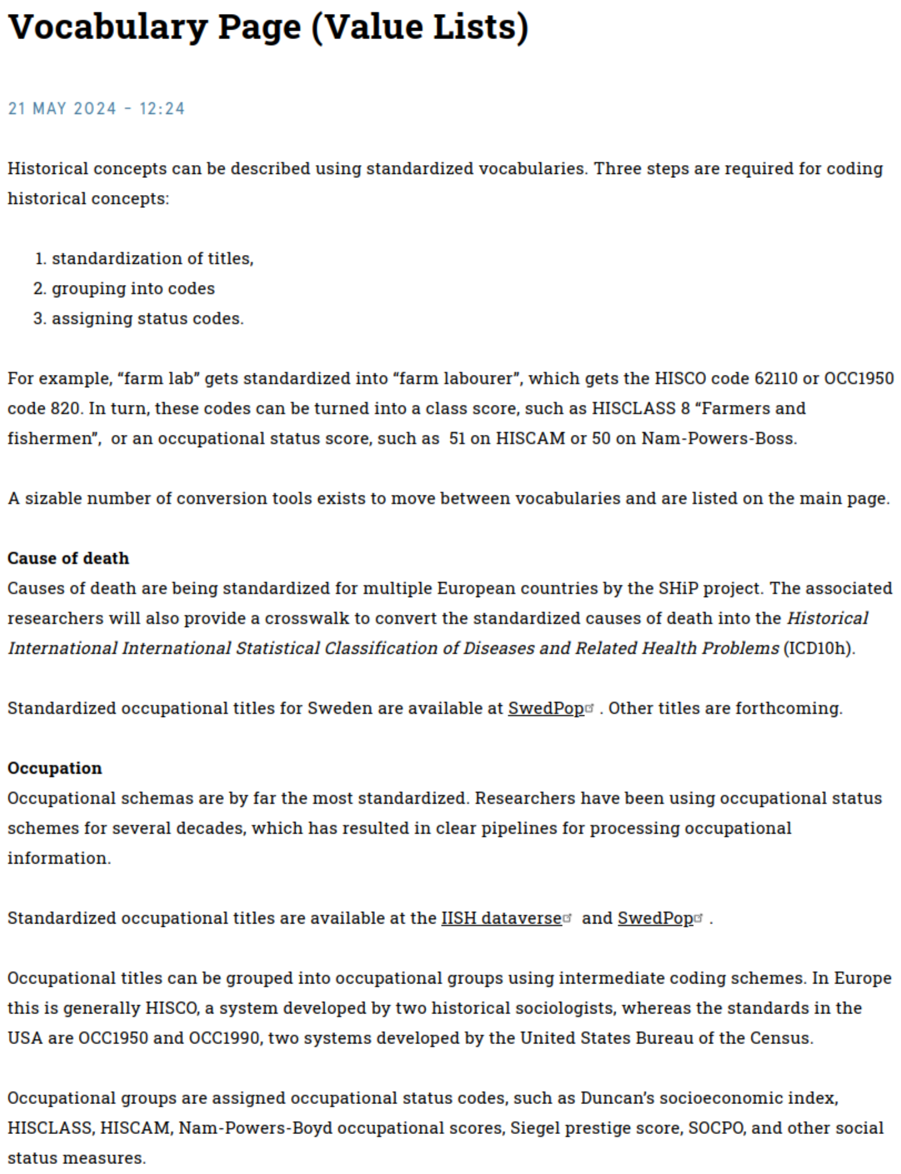
There are two kinds of schemas within historical demography. IPUMS-USA, IPUMS-International, MOSAIC, and NAPP were developed to standardize census data. IDS, LINKS-gen, and PiCo are efforts to standardize historical data from other types of historical sources, such as the civil registry, militia registers, parish registers, population registers, slave registers, or tax registers. All schemas cater to a very specific public, as they have been developed for projects with strong institutional boundaries.

Census schemas (IPUMS, MOSAIC, NAPP)

There are three schemas to describe census data. The driving force behind processing census data is IPUMS in Minnesota. In 1991, they started providing "common-format extracts" with common codes and constructed variables. In 1999, IPUMS joined up with scholars from Canada, Denmark, Great Britain, Iceland, Norway, Scotland in the North Atlantic Population Project (NAPP). A similar international census comparison project took place in the early 2010s at the Max Planck Institute for Demographic Research (MPIDR) in Rostock, Germany. The MPIDR schema, called MOSAIC, standardized census data from the 1700s until 1950 for 18 regions in Europe.

Theme	Variables	IPUMS-International	MOSAIC	NAPP
Geography	Country code Place Region Urban-rural status	CNTRY - - URBAN	country place region urban	CNTRY - - URBAN
Household	Group quarter status Household size Household weight	GQ PERSONS WTHH	gq hhsize hhwt	GQ NUMBERHH HHWT
Identifier	Enumeration Household Person	SAMPLE SERIAL PERNUM	id_enum id_hhold id_pers	SAMPLE SERIAL PERNUM

Figure 3: CLAIR-HD schema page



Vocabulary Page (Value Lists)

21 MAY 2024 - 12:24

Historical concepts can be described using standardized vocabularies. Three steps are required for coding historical concepts:

1. standardization of titles,
2. grouping into codes
3. assigning status codes.

For example, “farm lab” gets standardized into “farm labourer”, which gets the HISCO code 62110 or OCC1950 code 820. In turn, these codes can be turned into a class score, such as HISCLASS 8 “Farmers and fishermen”, or an occupational status score, such as 51 on HISCAM or 50 on Nam-Powers-Boss.

A sizable number of conversion tools exists to move between vocabularies and are listed on the main page.

Cause of death
Causes of death are being standardized for multiple European countries by the SHiP project. The associated researchers will also provide a crosswalk to convert the standardized causes of death into the *Historical International International Statistical Classification of Diseases and Related Health Problems* (ICD10h).

Standardized occupational titles for Sweden are available at SwedPop . Other titles are forthcoming.

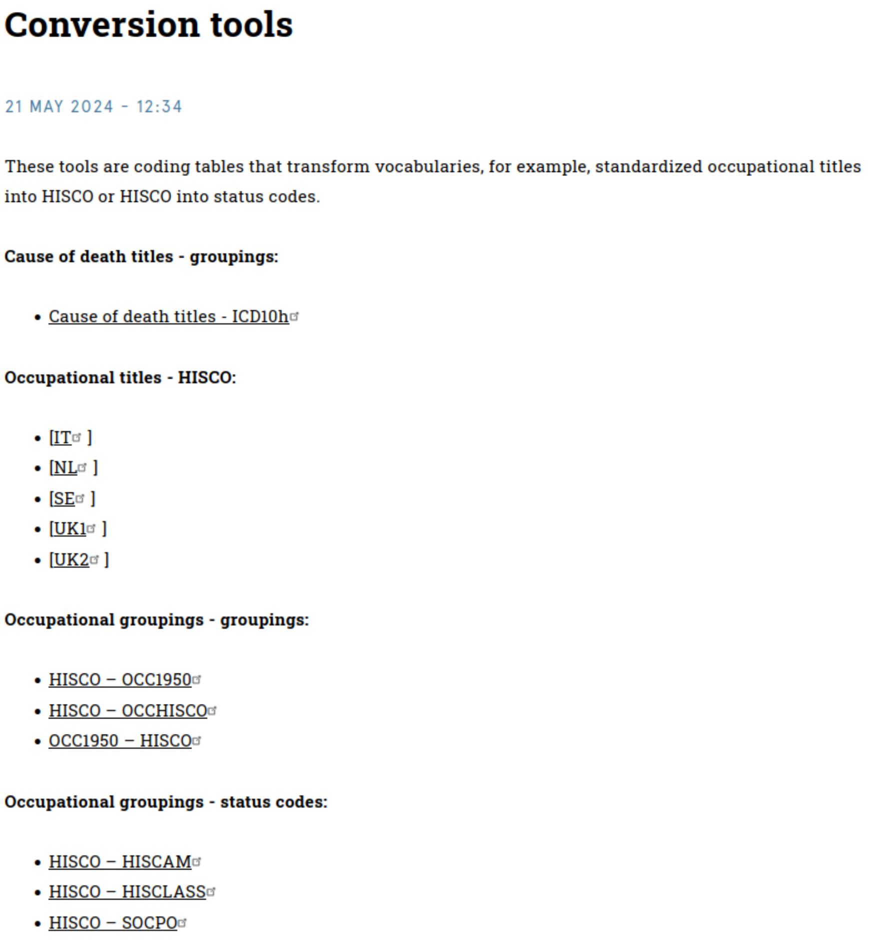
Occupation
Occupational schemas are by far the most standardized. Researchers have been using occupational status schemes for several decades, which has resulted in clear pipelines for processing occupational information.

Standardized occupational titles are available at the IISH dataverse and SwedPop .

Occupational titles can be grouped into occupational groups using intermediate coding schemes. In Europe this is generally HISCO, a system developed by two historical sociologists, whereas the standards in the USA are OCC1950 and OCC1990, two systems developed by the United States Bureau of the Census.

Occupational groups are assigned occupational status codes, such as Duncan’s socioeconomic index, HISCLASS, HISCAM, Nam-Powers-Boyd occupational scores, Siegel prestige score, SOCPO, and other social status measures.

Figure 4: CLAIR-HD vocabulary page



Conversion tools

21 MAY 2024 - 12:34

These tools are coding tables that transform vocabularies, for example, standardized occupational titles into HISCO or HISCO into status codes.

Cause of death titles - groupings:

- Cause of death titles - ICD10h

Occupational titles - HISCO:

- [IT]
- [NL]
- [SE]
- [UK1]
- [UK2]

Occupational groupings - groupings:

- HISCO – OCC1950
- HISCO – OCCHISCO
- OCC1950 – HISCO

Occupational groupings - status codes:

- HISCO – HISCAM
- HISCO – HISCLASS
- HISCO – SOCPO

Figure 5: CLAIR-HD conversion tool page

5 Conclusion & Discussion

There are multiple efforts within the field of historical demography to standardize information using schemas and vocabularies. There are two ways in which the field has developed standardization efforts. On the one hand, there are data centers that developed their own practices in how to standardize data. This information is of great use to other scholars, but only if these "habits" have become accepted practice in the field. Therefore, institutions have worked together on improving the interoperability of historical demographic data. However, the schemas that resulted from these collaborations have limited interoperability and are not widely adapted. There is considerable overlap between the concepts that schemas in historical describe. Yet, most schemas seem to be limited by the institutional context in which they were developed, as they describe concepts with their own terminology, rather than re-using existing vocabularies.

A notable exception is Persons in Context, which uses a concentric model to describe data to a wide an audience as possible [5]. A first general layer is used to to describe common concepts, domain-specific layer specifies terminology that are well-known within fields, and specialized vocabularies to standardize concepts unique to historical demography. This practice makes data more easily findable, accessible, and interoperable, especially smaller datasets and other "long-tail data" that can easily be obscured from view [57]. Yet, it is currently impossible to effectively describe historical datasets concentrically, due to a lack of specialized vocabularies within historical demography. Therefore, more specialized vocabularies are necessary to make historical demographic datasets (re)usable for a wider audience.

Specialized vocabularies within historical demography are mostly focused on occupational status on occupational titles, groupings, and status. This is indicative of the general process in the field where information is standardized, categorized, and operationalized. Currently, a network of researchers is working hard to standardize historical causes of death titles and simultaneously introduce a historical cause of death classification system. Yet, standardization of other historical concepts, such as place names, religious denominations, and data quality flags is lacking behind. SHiP, the project behind the standardization of cause of death information, shows that the development and adaptation of these specialized vocabularies can succeed by building on existing networks and adapting the principles of team science.

Surprisingly, the development of shared vocabularies to standardize variables is a relatively new phenomenon within the field. Systems to codify occupational clusters or determine occupational status are made by sociologists, rather than historians. The adaptation and use of these vocabularies shows that historical demographers are willing to use standardized variables and that they prove results. The efforts by the SHiP network [17] to standardize historical causes of death titles and codify them is the first example of historians working together to develop their own vocabularies and serves as a blueprint for how the field of historical demography, but also other fields within the humanities and social science, can develop new specialized vocabularies.

CLAIR-HD highlights the hard work done by other scholars. It gives an overview of what has been done, and what work could still be done. The website will be kept up to date for at least the coming 10 years, so that researchers and database managers can publish their data with similar standards or explain why existing standards are insufficient. The information provided by CLAIR-HD facilitates creative discussions and makes enquiries into historical demography easier and more insightful. By doing so, CLAIR-HD will serve as a case study that offers direction for other fields in the humanities and social sciences.

Acknowledgements This publication is part of the project "Common Language for Accessibility, Interoperability, and Reusability in Historical Demography" (CLAIR-HD) with project number 203.001.166 of the research programme Open Science Fund which is (partly) financed by the Dutch Research Council (NWO).

References

1. Alter, G., Mandemakers, K.: The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. Historical Life Course Studies **1**(1), 1-26 (2014).
2. Alter, G., Mandemakers, K., Gutmann M.: Defining and distributing longitudinal historical data in a general way through an intermediate structure. Historical Social Research **34**(3), 78-114 (2009).
3. Askhpour, A.: Linked International Classification for Religions. IISH Data Collection (2017). https://hdl.handle.net/10622/MHJWRZ
4. Basten, S.: Basten_Northern_English_Parishes_1777-1812, IISH Data Collection (2016). https://hdl.handle.net/10622/YK84PG
5. CBG.: Persons in Context. GitHub (2023) https://github.com/CBG-Centrum-voor-familiegeschiedenis/PiCo
6. Davis, I., Galbraith, D.: BIO: A vocabulary for biographical information. vocab.org (2004). https://vocab.org/bio
7. Duncan, O.D.: A socioeconomic index for all occupations, in A. Reiss, O.D. Duncan, P.K. Hatt, D.C. North (Eds.): Occupations and Social Status. New York: Free Press (1961).
8. Edvinsson, S., Mandemakers, K., Smith, K.R., Puschmann, P.: Harvesting: The result and impact of research based on historical longitudinal databases. Nijmegen: Radboud University Press (2023).
9. Fauve-Chamoux, A., Bolovan, I., Sogner, S.: A global history of historical demography: Half a century of interdisciplinarity. Bern: Peter Lang (2016).
10. Fornasin, A., Marzona, A.: HISCO_Italian_Formasin_Marzona_2006, IISH Data Collection (2016). https://hdl.handle.net/10622/SRVW6S
11. Gao, S.S., Sperberg-McQueen, C.M., Thompson, H.: W3C XML schema definition language (XSD) 1.1 part 1: Structures (2012) https://www.w3.org/TR/2012/REC-xmlschema11-1-20120405/
12. GeoNames Geographical Database. https://www.geonames.org/
13. Getty Thesaurus of Geographic Names® Online (2017). http://www.getty.edu/research/tools/vocabularies/tgn/
14. Grossner, K., Mostern, R.: Linked Places in World Historical Gazetteer, 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities (GeoHumanities'21), Beijing, China (2021). https://doi.org/10.1145/3486187.3490203

15. Guha, R.V., Brickley, D., Macbeth, S.: Schema.org: Evolution of structured data on the web. Communications of the ACM, 59(2), 44-51 (2016). https://doi.org/10.1145/2844544
16. Huijsmans, D.P.: HSN Gazetteer, IISH Data Collection (2020). https://hdl.handle.net/10622/ZDT2DJ
17. Janssens, A.: Constructing SHiP and an international historical coding system for causes of death. Historical Life Course Studies **10**, 64–70 (2021). https://doi.org/10.51964/hlcs9569
18. Janssens, A., Devos, I.; The limits and possibilities of cause of death categorisation for understanding late nineteenth century mortality. Social History of Medicine, 35**4**, 1053–1063 (2022). https://doi.org/10.1093/shm/hkac040
19. Lambert, P.S., Zijdeman, R.L., Van Leeuwen, M.H.D., Maas, I., Prandy, K.: The construction of HISCAM: A stratification scale based on social interactions for historical comparative research. Historical Methods: A Journal of Quantitative and Interdisciplinary History, **46**2, 77–89 (2013).
20. Mandemakers, K. et al.: HSNDB Occupations. IISH Data Collection (2020). https://hdl.handle.net/10622/88ZXD8
21. Mandemakers, K. (2023). “You really got me”. Ontwikkeling en toekomst van historische databestanden met microdata [Development and future of historical databases with microdata]. Rotterdam: Erasmus University Rotterdam. https://doi.org/10.25397/eur.23256467
22. Mandemakers, K., Alter, G., Vézina, H., Puschmann, P.: Sowing: The construction of historical longitudinal population databases. Nijmegen: Radboud University Press (2023).
23. Meroño-Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, K.S., Van Harmelen, F.: Semantic technologies for historical research: A survey. Semantic Web, 6**6**, 539–564 (2015). https://doi.org/10.3233/SW-140158
24. Meroño-Peñuela, A., De Boer, V., Van Erp, M., Zijdeman, R.L., Mourits, R.J., Melder, W., Rijpma, A., Schalk, R.: CLARIAH: Enabling interoperability between humanities disciplines with ontologies. In G. Cota, M. Daquino, G.L. Pozzato (eds.) Applications and Practices in Ontology Design, Extraction, and Reasoning. https://doi.org/10.3233/SSW200036
25. Minnesota Population Center: Integrated public use microdata series, International: Version 7.3. Minneapolis: IPUMS (2020). https://doi.org/10.18128/D020.V7.3
26. Mooney, G.: Mooney_1866_London_occupational_codes, IISH Data Collection (2016). https://hdl.handle.net/10622/ERGY0V
27. Mourits, R.J.: HISCO-OCC1950 crosswalk. DANS Easy (2017). https://doi.org/10.17026/dans-zap-qxmc
28. Mourits, R.J., Mandemakers, K., Laan, F., Munnik, C., Meijer, K.: HSNDB Standardisation Tables. IISH Data Collection (2024). https://hdl.handle.net/10622/IKB8HO
29. Mourits, R.J., Van Dijk, I.K., Mandemakers, K.: From matched certificates to related persons, **9**, 49–68 (2020). https://doi.org/10.51964/hlcs9310
30. Nam, C.B., Powers, M.G.: Changes in the relative status of workers in the United States, 1950-1960, Social Forces, 47, 158–170 (1968). https://doi.org/10.1093/sf/47.2.158
31. Nam, C.B., Boyd, M.: Occupational status in 2000: Over a century of census-based measurement, Population Research and Policy Review, 23, 327–358 (2004). https://doi.org/10.1023/B:POPU.0000040045.51228.34

32. Pedersen, B., Holsbø, E., Andersen, T., Shvetsov, N., Ravn, J., Sommerseth, H.L., Bongo, L.A.: Lessons learned developing and using a machine learning model to automatically transcribe 2.3 million handwritten occupation codes. arXiv preprint arXiv:2106.03996 (2020)
33. Quaranta, L., Sommerseth, H.L.: Introduction: Intergenerational Transmissions of Infant Mortality using the Intermediate Data Structure (IDS). Historical Life Course Studies **7**, 1-–10 (2018). https://doi.org/10.51964/hlcs9288
34. Reid, A., Garrett, E.: Doctors and the causes of neonatal death in Scotland in the second half of the nineteenth century. Annales de démographie historique, 123, 149–179 (2012). https://doi.org/10.3917/adh.123.0149
35. Revuelta-Eugercios, B., Castenbrandt, H., Løkke, A.: Older rationales and other challenges in handling causes of death in historical individual-level databases: The case of Copenhagen, 1880–1881. Social History of Medicine, 35**4**, 1116–1139 (2022). https://doi.org/10.1093/shm/hkab037
36. Roberts, E., Ruggles, S., Dillon, L. Y., Gardarsdottir, Ó., Oldervoll, J., Thorvaldsen, G., Woollard, M.: The North Atlantic Population Project an overview. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 36**2**, 80–88 (2003). https://doi.org/10.1080/01615440309601217
37. Roberts, E., Woollard, M., Ronnander, C., Dillon, L.Y., Thorvaldsen, G.: Occupational classification in the North Atlantic population project. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 36**2**, 89–96 (2003). https://doi.org/10.1080/01615440309601218
38. Ruggles, S., Roberts, E., Sarkar, S., Sobek, M.. The North Atlantic population project: Progress and prospects. Historical methods: A Journal of Quantitative and Interdisciplinary History, 44**1**, 1–6 (2011). https://doi.org/10.1080/01615440309601217
39. Ruggles, S.: The Minnesota Population Center data integration projects: Challenges of harmonizing census microdata across time and place. 2005 Proceedings of the American Statistical Association (2006), **9**, 1405–1415 (2006).
40. Ruggles, S., Fitch, C.A., Goeken, R., Hacker, J.D., Nelson, M.A., Roberts, E., Schouweiler, M., Sobek, M.: IPUMS ancestry full count data: Version 3.0. Minneapolis: IPUMS (2021). https://doi.org/10.18128/D014.V3.0
41. Siegel, P.M.: Prestige in the American occupational structure. Doctoral dissertation, University of Chicago (1971).
42. Song, X., Campbell, C.D.: Genealogical microdata and their significance for social science, Annual Review of Sociology, 43, 75–99 (2017). https://doi.org/10.1146/annurev-soc-073014-112157
43. Stapel, R.J.: 'Conflating Historical Population Statistics Using a Historical GIS with a Flexible Semantic Model for Premodern Administrative Units in the Low Countries: The *(Re)counting the Uncounted* and *Historical Atlas of the Low Countries* Projects', GeoHumanities '23: Proceedings of the 7th ACM SIGSPATIAL International Workshop on Geospatial Humanities, 56–59 (2023). https://doi.org/10.1145/3615887.3627756
44. Stapel, R.J.: 'Historical Atlas of the Low Countries. A GIS Dataset of Locality-Level Boundaries (1350–1800)', Research Data Journal for the Humanities and Social Sciences, 8, 1–32 (2023). https://doi.org/10.1163/24523666-bja10033
45. SwedPop: Swedish Population Databases for Research (January 2024). https://swedpop.se/
46. Szołtysek, M., Gruber, S.: Mosaic: Recovering surviving census records and reconstructing the familial history of Europe. The History of the Family, **21**1, 38–60 (2016). https://doi.org/10.1080/1081602X.2015.1006655

47. Van Erp, M., Tullett, W., Christlein, V., Ehrhart, T., Hürriyetoğlu, A., Leemans, I., Lisena, P., Menini, S., Schwabe, D., Tonelli, S., Troncy, R., Zinnen, M.; 'More than the name of the rose: How to make computers read, see, and organize smells.' The American Historical Review, 128**1**, 2023. https://doi.org/10.1093/ahr/rhad141
48. Van Herck, S., Mourits, R.J.: Upcycling the Dutch civil registry using Linked Data, legacy4reuse, Bamberg, Germany (2023).
49. Van Leeuwen, M.H.D., Maas, I., Miles, A.: HISCO: Historical international standard classification of occupations. Leuven University Press, Leuven (2002).
50. Van Leeuwen, M.H.D., Maas, I.: HISCLASS: A historical international social class scheme. Leuven University Press, Leuven (2011).
51. Sang-Ajang, H., Altink, N., Dikland, P, Jonkers, C., Kariomengolo, Valies, C., Van Oort, T.: Paramaribo Ward Registers 1828-1847. IISH Data Collection (2024). https://hdl.handle.net/10622/VLN8FD
52. Van der Meer, A., Boonstra, O.: Repertorium van Nederlandse gemeenten vanaf 1812 waaraan toegevoegd de Amsterdamse code. Den Haag: DANS (2011).
53. Van de Putte, B., Svensson, P.: Measuring social structure in a rural context Applying the SOCPO scheme to Scania, Sweden (17 (th)-20 (th) century). Belgisch Tijdschrift voor Nieuwste Geschiedenis-Revue Belge d'Histoire Contemporain, 40, 1-2, 249–293 (2010).
54. Van Zundert, J.: If you build it, will we come? Large scale digital infrastructures as a dead end for digital humanities, Historical Social Research/Historische Sozialforschung, 37**3**, 165–186 (2012). https://www.jstor.org/stable/41636603
55. US Bureau of the Census (1950). Alphabetical index of occupations and industries: 1950. Washington, D.C.: US Bureau of the Census.
56. WikiData. https://www.wikidata.org/
57. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific data, 3**1**, 1-9 (2016). https://doi.org/10.1038/sdata.2016.18
58. World Wide Web Consortium: Prov-o: The prov ontology (2013). https://www.w3.org/TR/prov-o/
59. Zijdeman, R.L.: OCCHISCO to HISCO. (2013). https://github.com/rlzijdeman/o-clack/tree/master/crosswalks/occhisco_to_hisco