Validating and Extracting Information from National Identification Numbers in R: the case of Finland and Sweden

by Pyry Kantanen, Erik Bulow, Måns Magnusson, Jussi Paananen, Leo Lahti

Abstract National identification numbers (NIN) are a widely used method for uniquely identifying individuals and companies in Finland, Sweden, and many other countries. Their widespread use in both public sector registers and private sector customer data systems prompts the need for openly available methods and tools for NIN analysis and validation. The **hetu** and **sweidnumbr** R packages provide functions for extracting information from Finnish and Swedish NINs. The packages can also generate temporary but structurally valid random numbers for testing purposes. Finnish and Swedish NINs contain information about person's birth date and sex, an individual number and a control character that helps detect most common input errors and structural irregularities. Our packages contain comprehensive documentation and examples to make usage of the package easy for domain experts with non-technical backgrounds. While the introduced tools are relevant mainly in the Finnish and Swedish context, this work contributes to developing similar tools for other countries. From a wider perspective, these packages contribute to the growing toolkit of standardized methods for computational social science research, epidemiology and other register-based inquiries.

Introduction

The **hetu** and **sweidnumbr** R packages provide open tools for handling and extracting data from identification codes for natural persons and juridical persons in the context of Finland and Sweden. Identification codes for natural persons are called *national identification numbers* (NIN) throughout this manuscript. NINs in the Nordic countries, a group which Finland and Sweden are a part of, contain embedded information that can be extracted and analysed.

Technical systems for identifying people, organizations, places and other objects are an important but often overlooked aspect of governance and management tools in modern societies (Dodge and Kitchin, 2005). Universal and persistent identification numbering systems for natural persons are vital facilitating research activities that combine data from different sources, for example in the fields of epidemiology, population studies and social research (Gissler and Haukka, 2004). Outside the field of academic research, universal identifiers for natural persons enables work in a multi-disciplinary and multi-agency contexts due to greater *administrative fluency* and *bureaucratic effectiveness* (Alastalo and Helén, 2022). This may be useful for example in the case of tackling wicked social problems that require co-operation from professionals from various fields, such as social workers, psychologists, police and health care professionals.

Prior R packages with similar scope include **numbersBR** for Brazilian identity numbers for individuals, vehicles and organisations (Freitas, 2018), and **generator** for generating various types of Personally Identifiable Information (PII), such as fake e-mail addresses, names and United States Social Security Numbers (Hendricks, 2015).

NIN generators and validators are therefore not a novel concept. In fact, in the case of Finland, handling Finnish NINs seems to be a common entry-level task to familiarize new computer science students with regular expressions, dates, string subsetting and similar concepts. The reasoning behind packaging these functions as R packages are similar to the ones listed in Wickham and Bryan's R Packages Introduction: code sharing, transparency, user manuals, semantic versioning and documenting changes between versions (Wickham and Bryan, 2022). In addition to these generally accepted virtues of open source software, there is of course intrinsic value in providing a solution to a common problem in the R language.

Features of identification number systems

Reliably keeping track of individuals, organizations, objects and flows in a given territory has long been seen as an important feature of modern governance (Dodge and Kitchin, 2005). Foucault (2009, 115-120) observed a historical pattern where practices first implemented in disciplinary institutions, such as prisons, military units and schools have spread to influence whole societies. We can see the results of this development today in the form of different identification code systems implemented around the world. Differences in legal, political and historical frameworks in different countries

have affected how these systems are implemented in practice, causing heterogeneity for example in identification system designs across Europe (Otjacques et al., 2007).

This heterogeneity as well as linguistic differences seem to contribute to variance in terminology used when referring to identification code systems. *Unique identifier* (UID) is an umbrella term that can be used to refer to unique identifiers for all sorts of things, from books (ISBN), chemicals (CAS) and legal entities (LEI) to anything imaginable (see Dodge and Kitchin, 2005). In this paper we are mainly interested in unique identifiers for natural persons and juridical persons.

Names such as *personal identification code* (Dodge and Kitchin, 2005), *personal identity number* (Alastalo and Helén, 2022) and *single identification number* (SIN) (Otjacques et al., 2007) are used in literature as a generic term. Personal identification codes can sometimes be confused with personal *personal identification numbers* (PIN) that refer to numeric or alphanumeric passcodes used for authentication, for example to withdraw cash or open a locked mobile phone. On the other hand names such as *personal identity code* (PIC) (Digital and Population Data Services Agency, 2022; Sund, 2012), name number (Watson, 2010) and personal number (Statistics Sweden, 2016) are used in official translations to refer to national implementations of NINs; in the mentioned cases, Finnish, Icelandic and Swedish NINs, respectively. Due to *function creep* (see Brensinger and Eyal, 2021; Alastalo and Helén, 2022) or *control creep* (see Dodge and Kitchin, 2005), historically sector-specific identifiers may also be used as a *de facto* NIN. This is the case with the US *social security number* (SSN) (Brensinger and Eyal, 2021) and Finnish employee pension card numbers and social security codes (Alastalo and Helén, 2022).

For clarity's sake, we will be using the generic term *national identification number* (NIN) throughout the manuscript to refer to all identification number systems and their implementations for natural persons, Finnish *personal identity codes* and Swedish *personal number* in particular. For organizations, we will use the generic term *organization identifier* when discussing Finnish *business IDs* (BID) and Swedish *organizational identity numbers* (OIN) / or Swedish organizational number (SON).

All identification code systems should strive to be both *unique* and *self-same* over time. Self-sameness refers to a degree of immutability that allows organizations to identify and reidentify a person over time. A combination of attributes[1] such as name, occupation and address would probably form a unique identifier even in relatively large crowds, but such attributes might not stay the same over time. (Brensinger and Eyal, 2021). In everyday life settings, a combination of personal attributes, such as name, occupation and address may be sufficiently unique to identify individuals even in large groups of people. Such an approach has its limitations. The problem with such composite keys is, however, that parts that make the composite key might not be permanent, making it impossible to compare different data points in the same dataset or combining data from different datasets.

According to Alterman (2003) a distinction can be made between *biocentric data* and *indexical data*. The former is biometric data connected to the physical features of the individual whereas the latter has no distinguishable relation to the individual, physiologically, psychologically, or otherwise. An example of biocentric data could be a fingerprint or an iris scan and an example of indexical data could be a randomly assigned number from which nothing can be deduced.[2]

For a number of reasons, many identification numbers are not just random strings. The American SSN originally contained information about the person's birth year and where the number was first registered (Brensinger and Eyal, 2021, 32) whereas Nordic countries' NINs often contain (or used to contain) information about the individual, usually birth date and sex (Watson, 2010; Salste, 2021). One reasoning for this was to make the code easier to remember (Alastalo and Helén, 2022). Even when sex and birth date are not biocentric data in the sense as (Alterman, 2003) defined it, including them takes Nordic NINs further away from being pure indexical data, thus making them more sensitive to handle. Table 1 provides a summary on the introduction of NINs in the Nordic countries as well as information which they contain.

In the Nordic countries comprehensive national identification number systems were developed and implemented from the 1940's to 1960's (Watson, 2010). In Sweden, the personal identity number (PIN) was introduced in 1947 as a way to identify individuals for tax purposes. The personal identity number consisted both of the date of birth and an additional three-digit birth number that together uniquely identified every individual in Sweden. In 1967 a check digit was added to easier identify incorrect entries to finalize the full Swedish personal identity number (or *personnummer*) (Åke

[1] Think of SQL's *composite key*

[2] Brensinger and Eyal (2021) discuss the concept of *dividuals*, manufactured objects that represent the living individual: address, fingerprints, name and so on. These dividuals need to go through the process of disembedding, standardization and reembedding to be useful. Disembedding means data gathering (taking a fingerprint sample), standardization means making the disembedded transcription into a standardized digital sample that can be easily compared with other similar samples and reembedding means linking these standardized records back to their actual flesh-and-blood counterparts. Without a way to reembed a huge and well standardized archive of fingerprints back to the population it is essentially useless. This is also a reason why biometric samples such as iris scans or fingerprints can never replace a primary keys in databases.

country	NIN name	introduced	characters (n)	birth date	sex	birth place
Sweden	personnummer	1947	11	yes	yes	yes
Iceland	kennitala	1950	10	yes	no	no
Norway	fødselsnummer	1964	11	yes	yes	no
Denmark	CPR-nummer	1968	11	yes	yes	no
Finland	henkilötunnus	1968	11	yes	yes	no

Table 1: Nordic NINs: year introduced and embedded information.

Johansson, 2003; Statistics Sweden, 2016).

The Finnish personal identity code has its roots in specialized employment pension number from 1962, which was gradually expanded to cover the whole population in the form of a social security number in 1964-1968. The structure of employment pension number was designed by a Finnish mathematician Erkki Pale, who had worked in Sweden in 1948-1951 and was most likely inspired by the Swedish regional birth numbers (Alastalo and Helén, 2022). His contribution was to design a more robust control character system to mitigate for input errors, especially in punch card systems. The design has proved to be resilient and with some minor tweaks it continues to be used, with the modern iteration being called a *personal identity code* (In Finnish: henkilötunnus, or *hetu* for short, hence the name of the package) (Salste, 2021). Similar to Finland, other Nordic countries took inspiration from Sweden as well (Krogness, 2011). Table 2 illustrates the similar structures of NINs in different Nordic countries.

country	NIN name	NIN example	NIN structure
SE	personnummer	610321-3499	YYMMDDCNNNQ
IS	kennitala	121212-1239	DDMMYYNNQC
NO	fødselsnummer	110779 41012	DDMMYYNNNQQ
DK	CPR-nummer	300280-1178	DDMMYY-NNNQ
FI	henkilötunnus	131052-308T	DDMMYYCNNNQ

Table 2: Examples of national identification numbers and their composition in five Nordic countries. DD: day, MM: month, YY: year, C: century marker, N: personal number numerical digit, Q: check digit or a control character.

In Finland the expansion of sector specific social security numbers and employment pension numbers to universal NINs in 1969 has contributed to the widespread use of secondary data sources[3] in research. One example of this development is the use of Finnish Hospital Discharge Register (FHDR), which contains data on all inpatient hospital discharges since 1967. According to Sund (2012) the proportion of erroneously inputed NINs or incomplete pseudo-NINs was initially high but fell rapidly as hospitals were specifically instructed to include it systematically and correctly to all discharge records starting from 1972. This has allowed to link discharge register data to other sources. In Sweden the PIN is currently used extensively in all parts of society, not only for taxation. It is used in education, for military service, in health care and by financial institutions, and insurance companies. The role of the Swedish NIN has also made it central to register-based research (Statistics Sweden, 2016).

Working with personal identity codes

The method of validating and extracting information from identification numbers is manually doable and simple in principle but in practice becomes unfeasible with datasets larger than a few dozen observations. The **hetu** and **sweidnumbr** packages provide easy-to-use tools for programmatic handling of Finnish and Swedish personal identity codes and Business ID codes.[4] As shown in Table 3, both packages share several core functions and even function names.

Both packages utilize R's efficient vectorized operations, generating and validating over 5 million personal identity codes or Business Identity Codes in less than 10 minutes on a regular laptop. This can meet the practical upper limit set by the current population of Finland (5.5 million people) and Sweden (10.35 million people), providing adequate headroom for handling of relatively large registry datasets.

[3]Secondary data: Data that has not been collected primarily for the purpose of a specific research question
[4]In Finnish: Yritys- ja Yhteisötunnus, or Y-tunnus for short, In Swedish: Organisationsnummer

sweidnumbr	hetu	Description
rpin	rpin (rhetu)	Generate a vector of random NINs
pin_age	pin_age (hetu_age)	Calculate age from NIN
luhn_algo	hetu_control_char	Calculate check digit / control character from NIN
pin_ctrl	pin_ctrl (hetu_ctrl)	Check NIN validity
pin_date (pin_to_date)	pin_date (hetu_date)	Extract Birth date from NIN
pin_sex	pin_sex (hetu_sex)	Extract Sex From NIN
oin_ctrl	bid_ctrl	Check OIN/BID validity
roin	rbid	Generate a vector of random OINs/BIDs

Table 3: Exported functions that are shared between both **sweidnumbr** and **hetu**. Function alias in parentheses.

The hetu package

Printing a data frame containing extracted information in a structured form can be done as follows:

```
> library(hetu)
> x <- c("010101A0101", "111111-111C", "290201A010M")
> hetu(x)
```

The `hetu()` function is the workhorse of the **hetu** package. Without additional parameters it prints out a `data.frame` with all information that can be extracted from Finnish NINs as well as a single column that indicates if the NIN is valid as a whole or if it has problems that make it invalid. For demonstration purposes the 3rd NIN we included has an invalid date part; 29th of February would only be a valid date if the year was a leap year, which 2001 is not.

```
         hetu    sex p.num ctrl.char       date day month year century valid.pin
1 010101A0101 Female   010         1 2001-01-01   1     1 2001       A      TRUE
2 111111-111C   Male   111         C 1911-11-11  11    11 1911       -      TRUE
3 290201A010M Female   010         M       <NA>  29     2 2001       A     FALSE
```

The generic way of outputing information found on individual columns is to use the standard `hetu()` function with extract-parameter. For example:

```
> hetu("010101A0101", extract = "sex")
[1] "Female"
> hetu("010101A0101", extract = "date")
[1] "2001-01-01"
```

All column names printed out by the `hetu()` function are valid extract parameters. Most commonly used columns have their own function wrappers that are identical in output:

```
> pin_sex("010101A0101")
[1] "Female"
> pin_date("010101A0101")
[1] "2001-01-01"
```

By importing **lubridate** (Grolemund and Wickham, 2011) functions we have also added `pin_age()` function to calculate the age of individuals in years (default), months, weeks or days. The default option is to calculate age at the current moment but it is also possible to set a specific date as a parameter at which the age is calculated.

```
> pin_age("010101A0101", date = "2004-02-01", timespan = "months")
The age in months has been calculated at 2004-02-01.
[1] 37
```

All NINs passed through the `hetu()` function are checked with 10 different tests to determine their validity. The results are crystallized in a single valid.pin column of the `hetu()` function output data frame. With the `hetu_diagnostic()` function the user can print the results of these normally hidden tests.

```
> hetu_diagnostic("290201A010M")
```

```
         hetu is.temp valid.p.num valid.ctrl.char correct.ctrl.char
1 290201A010M   FALSE        TRUE            TRUE             FALSE
  valid.date valid.day valid.month valid.year valid.length valid.century
1      FALSE      TRUE        TRUE       TRUE         TRUE          TRUE
```

When data is inputted manually without validity checks, input errors can creep in. The control character in Finnish personal identity codes combined with validity checks in **hetu** package can help to catch the most obvious errors. In the example above we can see that the date is incorrect, but also the control character is incorrect. We can simply try three different dates to see if the input error is in the date, month or year part, assuming that the personal number and control character parts were inputted correctly. In this manufactured example the error was in the year part, resulting in the rare leap day date being the correct one.

```
> example_vector <- c("290201A010M", "280201A010M", "290301A010M", "290200A010M")
> columns <- c("valid.p.num", "valid.ctrl.char", "correct.ctrl.char", "valid.date")
> hetu_diagnostic(example, extract = columns)

         hetu valid.p.num valid.ctrl.char correct.ctrl.char valid.date
1 290201A010M        TRUE            TRUE             FALSE      FALSE
2 280201A010M        TRUE            TRUE             FALSE       TRUE
3 290301A010M        TRUE            TRUE             FALSE       TRUE
4 290200A010M        TRUE            TRUE              TRUE       TRUE
```

The **hetu** package can generate a large number of personal identity codes with the `rpin` function. The date range of the generated identity codes can be changed with parameters, but it has a hard coded lower limit at the year 1860 and upper limit in the current date. It has been theorized that the oldest individuals that received a personal identity code in 1960s were born in 1860s. Personal identity codes are never assigned beforehand and therefore it is impossible to have valid personal identity codes that have a future date.

The function can also be used to generate so called temporary personal identity codes. Temporary identity codes are never used as a persistent and unique identifier for a single individual but as a placeholder in institutions such as hospitals when a person does not have a Finnish NIN or the NIN is not known. They can be identified by having a personal number (p.num column in `hetu()` function output or NNN as in Table 2) in the range of 900-999.

Here is an example of generating 4 temporary Finnish NINs and checking their validity with `pin_ctrl` function:

```
> set.seed(125)
> x <- rpin(n = 4, p.male = 0.25, p.temp = 1.0)
> x
[1] "250706-9565" "230117-940B" "291247-990E" "130271-908L"

> pin_ctrl(x)
[1] NA

> pin_ctrl(x, allow.temp = TRUE)
[1] TRUE TRUE TRUE TRUE
```

As mentioned earlier, our package also supports similarly generating and checking the validity of Finnish organization identifiers, or Finnish Business ID (BID) numbers. Despite the name, BIDs are used not only for companies and businesses but also for other types of organizations and other juridical persons. Unlike personal identity codes, BIDs do not contain any information about the company. BIDs consist of a random string of 7 numbers followed by a dash and 1 check digit, a number between 0 and 9.

In addition we have added support for the less known and less widely used numbering scheme for natural persons, Finnish Unique Identification (FINUID) numbers.[5] FINUID numbers are similar in the sense that they do not contain any biocentric data on the individual and are mainly used by government authorities in IT systems. FINUID numbers consist of 8 numbers and 1 control character calculated in the same way as in personal identity codes. Both of these ID numbers are an example of indexical data, as opposed to Finnish NINs which contain biocentric data in the form of birth date and sex.

```
> bid_ctrl(c("0000000-0", "0000001-9"))
[1] TRUE TRUE
```

[5] in Finnish: Sähköinen asiointitunniste or SATU for short

```
> satu_ctrl("10000001N")
[1] TRUE
```

The **hetu** package contains some functions that are not shared with the **sweidnumbr** package, most notable being `hetu()` function. These functions are listed and described in Table 4.

Function (alias)	Description
hetu	Finnish personal identification number extraction
pin_diagnostic (hetu_diagnostic)	Diagnostics Tool for HETU
satu_control_char	FINUID Number Control Character Calculator
satu_ctrl	Check FINUID Number validity

Table 4: Functions that are unique to the **hetu** package and have no equivalent in the **sweidnumbr** package. Function alias in parentheses.

The sweidnumbr package

The **sweidnumbr** R package has similar functionality as the **hetu** package, but for Swedish NINs and with a slightly different syntax. At the time of writing, the package has been downloaded roughly 30 000 times from CRAN[6]. The example NINs below taken from The Swedish Tax Agency (2007).

```
> library(sweidnumbr)
> example_pin <- c("640823-3234", "6408233234", "19640823-3230")
> example_pin <- as.pin(example_pin)
> example_pin

[1] "196408233234" "196408233234" "196408233230"
Personal identity number(s)
```

Unlike the **hetu** package, the **sweidnumbr** takes advantage of a custom S3 class structure. Therefore the first step is to convert strings with different Swedish NIN formats or numeric variables into a `pin` vector using the `as.pin()` function. The `pin` vector is a S3 object and can be checked is the `is.pin()` function.

```
> is.pin(example_pin)

[1] TRUE
```

This function only check that the vector is a `pin` object, but not if the actual NIN are valid. To check the Swedish NIN using the control numbers, or check digits, we simply use the `pin_ctrl()` function.

```
> pin_ctrl(example_pin)

[1]  TRUE  TRUE FALSE
```

Just as in the **hetu** package we can extract information from the Swedish NIN with specialized functions. We can now use `pin_birthplace()`, `pin_sex()`, and `pin_age()` to extract information on birthplace, sex, and age.

```
> pin_sex(example_pin)

[1] Male Male Male
Levels: Male

> pin_birthplace(example_pin)

[1] Gotlands län Gotlands län Gotlands län
28 Levels: Stockholm stad Stockholms län Uppsala län ... Born after 31 december 1989

> pin_age(example_pin)
[1] 55 55 55
```

[6]Source: CRANlogs API, data retrieved at 2022-03-22.

```
> pin_age(example_pin, date = "2000-01-01")
[1] 35 35 35
```

As with the **hetu** R package we can also generate, or simulate, NINs with the `rpin()` function. Shared functions exist also for Swedish organization identifiers, or Swedish organizational numbers (SON), in the form of `as.oin()`, `is.oin()`, and `oin_ctrl()` functions. Unlike the Finnish BID, the `oin` contain information on the type of organization of a given SON, which can be determined by using the `oin_group()` function.

```
> oin_group(example\_oin)

[1] Aktiebolag
[2] Stat, landsting, kommuner, församlingar
[3] Ideella föreningar och stiftelser
3 Levels: Aktiebolag ... Stat, landsting, kommuner, församlingar
```

Similar to **hetu** package `rbid()` function and **sweidnumbr** `rpin()` function for natural persons, we can generate new SONs using the `roin()` function.

```
> set.seed(125)
> roin(3)

[1] "776264-6144" "274657-0148" "827230-7631"
Organizational identity number(s)
```

Due to national characteristics of Swedish numbering schemes for natural and juridical persons there are some functions that are unique to the **sweidnumbr** packages. These functions are listed in Table 5.

Function	Description
as.oin	Parse organizational identity numbers
as.pin	Parse personal identity numbers to ABS format
fake_pins	Fake personal identity numbers and names
format_pin	Formatting pin
is.oin	Test if a character vector contains correct 'oin'
is.pin	Parse personal identity numbers to ABS format
oin_group	Calculate organization group from 'oin'
pin_birthplace	Calculate the birthplace of 'pin'
pin_coordn	Check if 'pin' is a coordination number

Table 5: Functions that are unique to the **sweidnumbr** package and have no equivalent in the **hetu** package.

Discussion

The **hetu** and **sweidnumbr** R packages provides a free and open source methods for validating and extracting data from a large number of Finnish and Swedish national identity numbers (NIN). While the package's target audience consists mostly of Finnish and Swedish users and people with particular interest in NIN systems around the world, the package makes a generic contribution in developing methodologies related to NIN handling in R, and more generally for *structured data* in the field of computational humanities (see e.g. (Mäkelä et al., 2020)). In the future, a more generic package or class structures could be useful to unify the handling of different NIN systems around the world, and might be worth exploring further.[7]

The origins of these packages can be traced to early 2010s when one curious individual wanted to analyze a large number of Finnish NINs that were leaked to the internet by an anonymous hacker. The legality and morality of handling such dataset containing personal information was and is in a grey area at best. As developers of this package we cannot condone such activities, even if they are

[7] Although in Unix philosophy one stated aim is "Write programs that do one thing and do it well" and "Write programs to work together" as opposed to writing one program that aims to do everything. The value in one generic program might mostly be related to sharing information and good coding practices among interested parties around the world.

conducted out of curiosity and not ill intentions, but we acknowledge that we cannot prevent our users from doing that either.[8]

We have acknowledged beforehand that random NINs generated with the **hetu** and **sweidnumbr** packages could theoretically be used for purposes such as synthetic identity fraud.[9] On the other hand it is important to note that such NINs could also be created by hand as information on valid NINs is readily available e.g. in the Finnish Digital and Population Data Services Agency and Swedish Tax Authority websites (Digital and Population Data Services Agency, 2022; The Swedish Tax Agency, 2007). Our package can be useful for many, and it does not make fraudulent activities significantly easier for malevolent individuals which is essential in judging the pros and cons of releasing this software to the public.

Similar data breaches have made people more wary of digital services. Privacy concerns can push Finland, Sweden and other Nordic countries towards redesigning their national identification numbers to omit some or all of the embedded personal information sometime in the future. For example in Finland there is at the time of writing an ongoing government project led by Ministry of Finance to redesign the NIN structure (Valtiovarainministeriö, 2022). There was no indication in literature that any country would be planning to join Germany and Hungary in banning universal identifiers for natural persons (see Otjacques et al., 2007).

Both packages are published under permissive BSD 2-clause license. We encourage our users to monitor for any legislative or policy changes related to NIN system implementations, give feedback and bug reports, study the source code and submit improvements to our public code repositories.[10]

Acknowledgements

We are grateful to all contributors, in particular Juuso Parkkinen and Joona Lehtomäki for their support in the initial package development. This work is part of rOpenGov[11] and contributes to the FIN-CLARIAH research infrastructure for computational humanities. LL and PK were supported by Academy of Finland (decisions 295741, 345630).

Bibliography

M. Alastalo and I. Helén. A code for care and control: The pin as an operator of interoperability in the nordic welfare state. *History of the Human Sciences*, 35(1):242–265, 2022. URL https://doi.org/10.1177/09526951211017731. [p1, 2, 3]

A. Alterman. "A piece of yourself": Ethical issues in biometric identification. *Ethics and information technology*, 5(3):139–150, 2003. ISSN 1388-1957. [p2]

J. Brensinger and G. Eyal. The Sociology of Personal Identification. *Sociological Theory*, 2021. URL https://doi.org/10.1177/07352751211055771. OnlineFirst. [p2, 8]

Digital and Population Data Services Agency. The personal identity code, 2022. URL https://dvv.fi/en/personal-identity-code. Accessed: 2022-01-17. [p2, 8]

M. Dodge and R. Kitchin. Codes of life: identification codes and the machine-readable world. *Environment and Planning D: Society and Space*, 23:851–881, 2005. [p1, 2]

M. Foucault. *Security, territory, population: lectures at the Collège de France, 1977-1978*. Palgrave Macmillan, New York, 2009. Editors: Michel Senellart, François Ewald, Alessandro Fontana, Arnold I. Davidson. [p1]

W. Freitas. numbersBR: Validate, Compare and Format Identification Numbers from Brazil, 2018. URL https://CRAN.R-project.org/package=numbersBR. R package version 0.0.2. [p1]

M. Gissler and J. Haukka. Finnish health and social welfare registers in epidemiological research. *Norsk Epidemiologi*, 14(1):113–120, 2004. [p1]

G. Grolemund and H. Wickham. Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25, 2011. URL https://www.jstatsoft.org/v40/i03/. [p4]

[8] The use of "Good, not evil" license texts is thought to be problematic and against the ethos of open source software

[9] see Brensinger and Eyal (2021, 32) for a short description of synthetic fraud related to American SSNs

[10] https://github.com/rOpenGov/hetu, https://github.com/rOpenGov/sweidnumbr

[11] http://ropengov.org

P. Hendricks. generator: Generate data containing fake personally identifiable information, 2015. URL https://CRAN.R-project.org/package=generator. R package version 0.1.0. [p1]

K. J. Krogness. Numbered individuals, digital traditions, and individual rights: civil status registration in Denmark 1645 to 2010. *Ritsumeikan Law Review*, 28:87–126, 2011. [p3]

E. Mäkelä, K. Lagus, L. Lahti, T. Säily, M. Tolonen, M. Hämäläinen, S. Kaislaniemi, and T. Nevalainen. Wrangling with non-standard data, 2020. [p7]

B. Otjacques, P. Hitzelberger, and F. Feltz. Interoperability of E-Government Information Systems: Issues of Identification and Data Sharing. *Journal of Management Information Systems*, 23(4):29–51, 2007. URL https://doi.org/10.2753/MIS0742-1222230403. [p2, 8]

T. Salste. Henkilötunnus – ihmisten koodaaja, 2021. URL https://www.tuomas.salste.net/doc/tunnus/henkilotunnus.html. Accessed: 2021-12-13. [p2, 3]

Statistics Sweden. Personal identity number, 2016. [p2, 3]

R. Sund. Quality of the Finnish Hospital Discharge Register: A systematic review. *Scandinavian journal of Public Health*, 40:505–15, 8 2012. doi: 10.1177/1403494812456637. [p2, 3]

The Swedish Tax Agency. Personnummer: Skv 704 ed. 8, 2007. [p6, 8]

Valtiovarainministeriö. Project on redesigning the system of personal identity codes, 2022. URL https://vm.fi/en/project-on-redesigning-the-system-of-personal-identity-codes. Accessed: 2022-01-17. [p8]

I. Watson. A short history of national identification numbering in Iceland. *Bifröst Journal of Social Science / Tímarit um félagsvísindi*, 1:51–89, 2010. ISSN 1670-7796. [p2]

H. Wickham and J. Bryan. R packages, 2022. URL https://r-pkgs.org/intro.html. The book is a work-in-progress. Accessed: 2022-01-17. [p1]

Åke Johansson. Från bläckpenna till datorhjärna. *Deklarationen 100 år och andra tillbakablickar*, 2003. [p2]

Pyry Kantanen
Department of Computing
PO Box 20014 University of Turku
Finland
ORCiD: 0000-0003-2853-2765
pyry.kantanen@utu.fi

Måns Magnusson
Department of Statistics
Uppsala University
Sweden
ORCiD: 0000-0002-0296-2719
mans.magnusson@statistik.uu.se

Leo Lahti
Department of Computing
PO Box 20014 University of Turku
Finland
ORCiD: 0000-0001-5537-637X
leo.lahti@utu.fi