

Validating and Extracting Information from National Identification Numbers in R: The Case of Finland and Sweden

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

Abstract National identification numbers (NIN) and similar identification code systems are widely used for uniquely identifying individuals and organizations in Finland, Sweden, and many other countries. To increase the general understanding of such techniques of identification, openly available methods and tools for NIN analysis and validation are needed. The `hetu` (Kantanen et al., 2022) and `sweidnumbr` (Magnusson and Bulow, 2020) R packages provide functions for extracting embedded information, checking the validity, and generating random but valid numbers in the context of Finnish and Swedish NINs and other identification codes. In this article, we demonstrate these functions from both packages and provide theoretical context and motivation on the importance of the subject matter. Our work contributes to the growing toolkit of standardized methods for computational social science research, epidemiology, demographic studies, and other register-based inquiries.

Introduction

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 for 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 enable work in a multi-disciplinary and multi-agency context 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 cooperation from professionals from multiple fields, such as social workers, psychologists, police, and health care professionals.

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 national contexts of Finland and Sweden. Prior R packages with similar scope include `numbersBR` for Brazilian identity numbers for individuals, vehicles, and organizations (Freitas, 2018), `cprr` for Danish "Det Centrale Personregister" (CPR) numbers (Anhøj, 2019), and `generator` for generating various types of Personally Identifiable Information (PII), such as fake e-mail addresses, names and the United States Social Security Numbers (Hendricks, 2015).

Identification code generators and validators are not a novel concept. In the case of Finland, the handling of identification codes can be seen as a common entry-level task for new computer science students to familiarize themselves with regular expressions, handling dates, string subsetting and similar concepts, and examples in various programming languages can easily be found online. However, the reasoning for packaging functions for Finnish and Swedish identification codes as yet another R package is the same as with any other R package: (1) code sharing, (2) transparency, (3) user manuals, (4) semantic versioning and (5) documenting changes between versions (as listed by Wickham and Bryan, 2022). In addition to the intrinsic values of open-source publishing and solving common problems with R, we aim to provide much-needed context to all this.

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 when alphanumeric identifiers originally assigned to prisoners, soldiers and students have been transformed into nationwide identification code systems 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 (Ottjacques et al., 2007).

This heterogeneity and linguistic differences seem to contribute to variance in the 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 generic terms. Personal identification codes can sometimes be confused with *personal identification numbers* (PINs) 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) to refer to all identification number systems and their country-specific implementations for natural persons, in this case, the Finnish *personal identity code* and the Swedish *personal number*. For organizations, we will use the generic term *organization identifier* when discussing Finnish *business IDs* (BID) and Swedish *organizational identity numbers* (OIN) / *Swedish organizational numbers* (SON).

All identification 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 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](#)).

According to [Alterman \(2003\)](#), a distinction can be made between *biocentric data* and *indexical data*. The former is biometric data connected to the individual's physical features 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.¹

For several 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 reason 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 arguably more sensitive to handle. Table 1 provides a summary of the introduction of NINs in the Nordic countries as well as information which they contain.

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.

In the Nordic countries, comprehensive national identification number systems were developed and implemented from the 1940s to the 1960s ([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 of both the date of birth and an additional three-digit birth number that together

¹[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 re-embedding to be useful. Disembedding means data gathering (e.g. 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 re-embedding means linking these standardized records back to their actual flesh-and-blood counterparts. Without a way to re-embed 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 primary keys in databases.

uniquely identified every citizen and permanent resident in Sweden. In 1967 a check digit was added to detect incorrect entries and thus the design of the Swedish personal identity number (or *personnummer*) was finalized (Åke Johansson, 2003; Statistics Sweden, 2016).

The Finnish personal identity code has its roots in specialized employment pension numbers 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 from 1948 to 1951 and was most likely inspired by the Swedish "regional birth number system" (Alastalo and Helén, 2022), most likely meaning the early iteration of the Swedish NIN. His contribution was to design a more robust control character system to mitigate input errors, originally in the context of 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*² (Salste, 2021).

Like Finland, other Nordic countries took inspiration from Sweden as well (Krogness, 2011). Table 2 illustrates the structural similarity of NINs in different Nordic countries. Some national variation does exist, however. In Norway and Denmark, individual numbers are used to denote the century in the birth date in addition to differentiating individuals from one another (Furseth and Ljones, 2015; CPR-kontoret, 2008). Since 2007 the Danish CPR numbers have dropped the check digit in favour of having an extra individual number to expand the pool of available unique numbers from roughly 500 per day to 4000-6000 per day (CPR-kontoret, 2008; Jerlach, 2009). All Nordic NIN designs prioritize a shorter length of 10-11 characters and use a combination of 2-digit years and a century marker or certain individual number ranges to denote the full year. For example, "52-" from the Finnish number and "99551" from the Norwegian number translate to 1952 and 1899, respectively. The more unambiguous Swedish 12-digit variant is used exclusively in automatic data processing systems (ADB, from Swedish *automatisk databehandling*) and not in day-to-day interactions.

country	NIN name	NIN example	NIN structure
Sweden	personnummer	610321-3499	YYMMDDCNNNQ
Iceland	kennitala	121212-1239	DDMMYYNNQC
Norway	fødselsnummer	01129955131	DDMMYYNNNQ
Denmark	CPR-nummer	300280-1178	DDMMYY-NNNN
Finland	henkilötunnus	131052-308T	DDMMYYCNNNQ
Sweden	personnummer (ADB)	196103213499	YYYYMMDDNNNQ

Table 2: Examples of national identification numbers and their composition in five Nordic countries. DD: day, MM: month, YY: year, C: century marker, N: individual number / serial number, 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 widespread secondary use of different data sources³ in public administration, education, and research and development. One example of this development is the use of the Finnish Hospital Discharge Register (FHDR), which contains data on all inpatient hospital discharges since 1967. The systematic development of record-keeping practices and staff training has made it possible to link discharge register data reliably to other data sources from 1972 onwards (Sund, 2012). In Sweden, the NIN 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). It could be argued that the most important feature of NIN systems is the interoperability it enables between different sectors of society (Alastalo and Helén, 2022).

Working with national identification numbers in R

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 codes⁴. 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

²In Finnish: henkilötunnus, or *hetu* for short, hence the name of the package

³Secondary data: Data that has not been collected primarily for a specific research question

⁴In 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.

personal identity codes or Business Identity Codes in less than 10 minutes on a regular laptop⁵. This can meet the practical upper limit set by the current population of Finland (5.5 million people) (Official Statistics of Finland (OSF), 2022) and Sweden (10.5 million people) (Statistiska centralbyrån, 2022), providing adequate headroom for the handling of relatively large registry datasets containing information on people currently alive and deceased.

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 any problems that make it invalid. For demonstration purposes the 3rd NIN listed below has an invalid date part; the 29th of February would only be a valid date if the year was a leap year, which we know that 2001 is not. The NIN would be correct if the year was changed from 2001 to 2000.

	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 full birth year is constructed by reading the 2-digit year information and the century marker; in the case of the first-row example, "01" and "A". "A" means that the person is born in the 2000s, "-" means the 1900s and "+" means the 1800s. A binary sex classification can be constructed simply by calculating if the personal number (p.num column) is an odd or an even number: Even numbers (for example "010") denote a female and odd numbers (for example "111") denote a male. The final character of the NIN is the control character, which is determined by dividing the concatenated integers (for example 290201010) by 31 and using the remainder as a key to retrieve a value from a list that includes numbers between 0 and 9 and English alphabets (without letters that might be mixed with numbers: I, G, O, Q and Z) (Digital and Population Data Services Agency, 2022; Salste, 2021).

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

```
> 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 wrapper function that are, as expected, identical in output:

```
> pin_sex("010101A0101")
[1] "Female"
```

⁵Tested on a 2015 Macbook Pro with Intel i5-5257U @ 2.70GHz

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

With the help of imported functions from the **lubridate** package (Grolemund and Wickham, 2011), we can calculate ages from NINs not only in years and days but also in months and weeks by using the `pin_age()` function. By default, the age is calculated in years at the current date and time but this end date can also be manually set by using the `date` parameter.

```
> 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. All tests need to be passed for a NIN to be valid. The results from different tests are summarized in the `valid.pin` column of the `hetu()` function output data frame. The user can print individual test results with the `hetu_diagnostic()` function for debugging purposes.

```
> 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 the `hetu()` function 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 incorrect⁶. We can simply try three different dates to see if the input error is in the day, 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 hardcoded lower limit at the year 1860 and an upper limit at the current date. It has been theorized that the oldest individuals that received a personal identity code in the 1960s were born in the 1850s or 1860s. Personal identity codes are never assigned beforehand and therefore it is impossible to have valid personal identity codes that have a future date. (Salste, 2021.)

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.

Below is an example of generating 4 temporary Finnish NINs and checking their validity with the `pin_ctrl()` function. Since all NINs are temporary, they do not pass the check validity checks meant for normal pins if they are not explicitly allowed. A vector with no valid NINs returns a single NA.

```
> 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)
```

⁶By validity we mean that the control character itself is an allowed character. By correctness, we mean that the inputted control character matches the calculated control character


```
[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.⁷ FINUID numbers consist of 8 random numbers and 1 control character that is calculated in the same way as in Finnish NINs. FINUID numbers are similar to BID numbers in the sense that they do not contain any biocentric data on the individual or the corporation⁸, but unlike BID numbers that are ubiquitous in corporate documents and public databases, FINUID numbers are mainly used by government authorities in internal IT systems.

```
> bid_ctrl(c("0000000-0", "0000001-9"))
```

```
[1] TRUE TRUE
```

```
> satu_ctrl("10000001N")
```

```
[1] TRUE
```

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

Function (alias)	Description
<code>hetu</code>	Finnish personal identification number extraction
<code>pin_diagnostic (hetu_diagnostic)</code>	Diagnostics Tool for HETU
<code>satu_control_char</code>	FINUID Number Control Character Calculator
<code>satu_ctrl</code>	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 CRAN⁹. The example NINs below are taken from the example published by the Swedish Tax Authority [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 `as.pin()` function formats all inputted numbers to a

⁷FINUID in Finnish: *sähköinen asiointitunniste* (SATU)

⁸Aside from sole trader/business name companies that are closely related to the individual entrepreneur, the term "biocentric" is badly suited when talking about corporations and other juridical persons. However, it could be argued that if the BID number contained information such as the company form, the place of registration or the date of registration it could be seen as analogous to biocentric information contained in NINs.

⁹Source: CRANlogs API, data retrieved at 2022-03-22.

so-called ADB-format¹⁰ with 12 digits and no century marker, which results in less ambiguity and no need to change the century marker from "-" to "+" when a person turns 100 years old. The `pin` vector is an S3 object and can be checked by using the `is.pin()` function.

```
> is.pin(example_pin)
```

```
[1] TRUE
```

This function only checks that the vector is a pin object, but not if the actual NINs 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 county of birth (for NINs assigned before 1990), 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
```

```
> 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` number contains 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 the `rbid()` function from the **hetu** package, we can generate new SONs using the `roin()` function from the **sweidnumbr** package.

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

```
> roin(3)
```

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

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

Discussion

The **hetu** and **sweidnumbr** R packages provide free and open-source methods for validating and extracting data from a large number of Finnish and Swedish national identity numbers (NIN). While

¹⁰ADB: Short from the Swedish term *automatisk databehandling*, meaning *automatic data processing* (ADP) in English

Function	Description
as.oin	Parse organizational identity numbers
as.pin	Parse personal identity numbers to ADP format
format_pin	Formatting pin
is.oin	Test if a character vector contains correct 'oin'
is.pin	Parse personal identity numbers to ADP 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.

the packages' target audience most likely mainly consists of Finnish and Swedish users and people with a particular interest in NIN systems around the world, the packages make a generic contribution to developing methodologies related to NIN handling in R, and more generally for *structured data* in the field of computational humanities (see Mäkelä et al., 2020), epidemiology and demographic studies (see Gissler and Haukka, 2004). A possible direction for future developments could be to create more generic class structures or even a completely new R package that could recognize and handle NIN systems from several different countries around the world.

The origins of the **hetu** package can be traced to the 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, to identify the source of the leak. The legality and morality of handling such datasets containing personal information was and is in a grey area at best. As developers of these packages, we cannot condone such activities, even if they are conducted out of curiosity and not of ill intentions, but we acknowledge that we cannot prevent our users from doing that either. Both **hetu** and **sweidnumbr** packages are free software with permissive licenses and pre-emptively limiting their use to only "good, not evil" causes would be problematic as well.¹¹

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.¹² 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. on 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 warier about 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, at the time of writing there is an ongoing government project led by the Finnish Ministry of Finance to redesign the Finnish NIN structure (Valtiovaraministeriö, 2022). We will continue to monitor for such policy changes and make changes to the packages if necessary.

As mentioned earlier, both packages are published under a permissive BSD 2-clause license. We encourage our users to give feedback on the packages and their materials, report bugs or any legislative or policy changes related to NIN system implementations, study the source code and submit improvements to our public code repositories¹³ or fork the code to better suit their needs.

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 rOpenGov¹⁴ and contributes to the FIN-CLARIAH research infrastructure for computational humanities. LL and PK were supported by the Academy of Finland (decisions 295741, 345630).

¹¹For example, JSON has a license that states that "The Software shall be used for Good, not Evil". Defining what is good and evil is at least in part up to everyone's personal judgment, making the license clause ambiguous.

¹²see Brensinger and Eyal (2021, 32) for a short description of synthetic fraud related to American SSNs

¹³<https://github.com/rOpenGov/hetu>, <https://github.com/rOpenGov/sweidnumbr>

¹⁴<http://ropengov.org>

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. Anhøj. cpr: Functions for Working with Danish CPR Numbers, 2019. URL <https://CRAN.R-project.org/package=cpr>. R package version 0.2.0. [p1]
- J. Brensinger and G. Eyal. The Sociology of Personal Identification. *Sociological Theory*, 2021. URL <https://doi.org/10.1177/07352751211055771>. OnlineFirst. [p2, 8]
- CPR-kontoret. Personnummeret i CPR-systemet, 2008. URL <https://cpr.dk/media/12066/personnummeret-i-cpr.pdf>. Accessed: 22.4.2022. [p3]
- Digital and Population Data Services Agency. The personal identity code, 2022. URL <https://dvv.fi/en/personal-identity-code>. Accessed: 2022-01-17. [p2, 4, 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]
- J. Furseth and O. Ljones. 50-årsjubilant med behov for oppgradering. *Samfunnsspeilet*, 2015(1), 2015. URL <https://www.ssb.no/befolkning/artikler-og-publikasjoner/50-arsjubilant-med-behov-for-oppgradering>. [p3]
- M. Gissler and J. Haukka. Finnish health and social welfare registers in epidemiological research. *Norsk Epidemiologi*, 14(1):113–120, 2004. [p1, 8]
- G. Grolemond 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/>. [p5]
- 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]
- T. Jerlach. Udviklingen på CPR-området i de seneste 20-25 år frem til 2009, April 2009. URL <https://cpr.dk/media/12060/udviklingen-paa-cpr-omraadet-frem-til-2009.pdf>. [p3]
- P. Kantanen, M. Magnusson, J. Paananen, and L. Lahti. hetu: Structural Handling of Finnish Personal Identity Codes, 2022. URL <https://github.com/rOpenGov/hetu>. R package version 1.0.7. [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]
- M. Magnusson and E. Bulow. sweidnumbr: R tools to handle of swedish identity numbers, 2020. URL <http://github.com/rOpenGov/sweidnumbr>. R package version 1.4.3. [p1]
- E. Mäkelä, K. Lagus, L. Lahti, T. Säily, M. Tolonen, M. Hämäläinen, S. Kaislaniemi, and T. Nevalainen. Wrangling with non-standard data, 2020. [p8]
- Official Statistics of Finland (OSF). Preliminary population statistics [online publication], March 2022. URL <https://www.stat.fi/en/publication/ckti21wgb3db0b531gwi04h8>. Accessed: 22.4.2022. [p4]
- 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>. [p1, 2]
- T. Salste. Henkilötunnus – ihmisten koodaaja, 2021. URL <https://www.tuomas.salste.net/doc/tunnus/henkilotunnus.html>. Accessed: 2021-12-13. [p2, 3, 4, 5]

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

Statistiska centralbyrån. SCB statistikdatabasen. [Elektronisk resurs] : Statistical database, 2022. URL <https://www.scb.se/hitta-statistik/statistik-efter-amne/befolkning/befolkningens-sammansattning/befolkningsstatistik/pong/tabell-och-diagram/manadsstatistik--riktet/befolkningsstatistik-2022/>. Accessed: 22.4.2022. [p4]

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. [p3]

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

Erik Bülow
Department of Orthopaedics, Institute of Clinical Sciences
Sahlgrenska Academy at University of Gothenburg
Sweden
ORCID: 0000-0002-9973-456X
erik.bulow@gu.se

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

Jussi Paananen
Institute of Biomedicine
University of Eastern Finland
Finland
ORCID: 0000-0001-5100-4907
jussi.paananen@uef.fi

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