

# Persons in Context: Potential and Pitfalls for the CLARIAH Infrastructure

Sytze Van Herck<sup>1,2</sup>, Ivo Zandhuis<sup>3</sup>, Richard L. Zijdemans<sup>3,4,5</sup>, and  
Rick J. Mourits<sup>3</sup>

<sup>1</sup>Utrecht University

<sup>2</sup>Ghent University

<sup>3</sup>International Institute of Social History

<sup>4</sup>VU Amsterdam

<sup>5</sup>University of Stirling

Historians and social scientists have been working with historical person data for decades, resulting in regional and national reconstructions of historical populations. However, current initiatives hardly allow for interlinking reconstruction projects, despite advances in computing power that allow population reconstructions of an ever-increasing size. With the Persons in Context (PiCo) vocabulary, the Center for Family History in the Netherlands (CBG) provides an opportunity to standardise historical person observations and interlink person observations from multiple sources into life courses in a FAIR way. As a result, PiCo could also facilitate modular person reconstruction software, allowing tools for cleaning, linking and evaluation to work in a single pipeline. The main aim of the paper is to describe the practical potential and predicaments of adopting PiCo in the current CLARIAH linked data pipeline for person reconstruction. In particular, we contribute (1) the incongruencies between the data model of PiCo and the CLARIAH record linkage tool, *burgerLinker*, and *C2RC*, the CLARIAH record linkage evaluation tool, (2) enhancing the interoperability of tools and data in a person reconstruction pipeline and (3) we present the concept of ‘provenance trails’, enabling a direct link between archival sources and research outcomes.

**Keywords:** [Data Models, FAIR, Person Observation, Person Reconstruction, PiCo-M, Provenance Trails, Research Infrastructure]

## 1 Introduction

The introduction of historical databases of person reconstructions 35 years ago initiated a new era for historical demography (Fauve-Chamoux et al., 2016). Archival sources containing personal demographic variables can be collected and organised in digital

resources to enhance our understanding of the past. The method of modeling the observations of personal characteristics is very important. Reconstructions of persons and family relations must follow standards to support combining or comparing datasets over time and space. Standards also ensure the quality and accountability of the data (Meroño Peñuela et al., 2020; Woltjer et al., 2024).

The importance of traceability is underlined by the long tradition in humanities research of making references to the sources on which the result is based. These references are traditionally organised in textual form as footnotes, end notes, or table captions. Such sources may include not only primary source material, but also secondary literature. There are important reasons to facilitate 'provenance trails' from archival source data to final research output. For one, references facilitate the reproducibility of the study and as a result increase its trustworthiness. References also reduce the time required to reproduce results, enhancing opportunities to evaluate research results. Finally, references explicate assumptions made in the data wrangling or research interpretation phase. This allows work to be revisited in the future, when more knowledge becomes available on assumptions made today.

Yet, in this day and age of Digital Humanities the field has hardly made progress in enhancing the degree of traceability. Reconstruction of historical persons requires data to flow from archival institutions to genealogists and researchers. Archival institutions scan, transcribe and disclose person observations from historical records. These observations are shared online for genealogists and researchers, who then create life course reconstructions based on these observations using manual methods or automated algorithms, of which the LINKS project has been the largest and most successful example in the Netherlands (Mandemakers et al., 2023). However, software made by different researchers and institutes is often not reusable, and the rationale behind reconstructions, also known as provenance, is often not retraceable. Moreover, in communicating the life course reconstructions, all links to underlying sources are aggregated into a handful of footnotes on which archival materials were used.

Having more detailed references to underlying sources, which we will call 'provenance trails', is feasible. Many of the larger research platforms stem from the 1990s and early 2000s, when FAIR was not yet an accepted philosophy, and especially memory or the use of the internet to link resources was limited. Yet, the same can be said for platforms such as *WieWasWie.nl* and *OpenArchieven.nl*, which both show how aggregator platforms can give credit to the original source in different ways, by directly pointing to that source online.

The fact that this state-of-the-art state is not yet embraced by research is in our view problematic for at least three reasons. The first is that the actual reproduction of research is still a very cumbersome process, requiring many manual, none traceable, actions. Second, as with research, archives are increasingly dependent on the explication of relevance to science and society, often expressed in usage indicators for funding. The occasional acknowledgment of archives in research papers is not retraceable for those archives and cannot be used to show their relevance to research. The third is a possibly waning relationship between GLAM institutions and academic research. FARO, the Flemish support institute for cultural heritage, has already called on Flemish heritage institutions to democratise their research function and emphasises the need to step away from measuring expertise based on academic and scientific merit releasing their exclusive authority as heritage institutions (Van Oost and Vander Stichele, 2024, p. 21).

In the Netherlands, there is a long-standing relationship between exchange of his-

torical person records via the archives, centralised by the Center for Family History (CBG), and person and family reconstruction via research embodied in the HSNDB at the International Institute of Social History (IISH). Research-made reconstitutions are returned to the CBG’s website as suggestions to help persons find their ancestors.<sup>1</sup> The CBG has recently, with the research and heritage community, released a Linked Data Vocabulary for the description of persons observations and person reconstructions in the heritage (archival) setting and the research setting. This vocabulary is called “Persons in Context” (PiCo) and will phase out an existing XML data model to support the ever-increasing exchange and use of historical person data, as meticulously described in Woltjer et al. (2024). The Netwerk Digitaal Erfgoed (NDE) urges Dutch GLAM institutions to provide their collections as linked data, archives are thus stimulated to use PiCo to describe person observations in their collection (Gaakeer et al. (2021); NDE and ministerie van OC&W (2024)). Linked Data “refers to a set of best practices for publishing structured data on the Web”.<sup>2</sup> Linked Data follows existing and widely adopted standards using vocabularies such as PiCo to model the data.

In this paper, we want to show how the PiCo vocabulary could be used to enhance the link between research and archival sources.<sup>3</sup> We will do so within the Common Lab Research Infrastructure for the Arts and Humanities (CLARIAH) infrastructure framework, and discuss the implications of Persons in Context for the CLARIAH person reconstruction and evaluation tools, burgerLinker and C2RC.<sup>4</sup> We also introduce the notion of ‘provenance trails’, a ‘bread crumb trail’ from research result to primary source as a means to capture the way that person reconstructions are built. By doing so, we underline the value of PiCo for (1) reusability of data and tooling, (2) enhancing the interoperability of tools and data in a person reconstruction pipeline and (3) the retraceability of research results to its archival resources. The importance of these concepts extends far beyond the presented pipeline and will be pivotal for sustainable Digital Heritage and Digital Humanities.

## 2 Generating Person Reconstructions

We first evaluate the software that generates person reconstructions from person observations in the indexed civil registry for the HSNDB project LINKS (Mandemakers et al., 2023). A Person Observation refers to a single instance of a person observed in a single source. LINKS aims to reconstruct all persons who were born, married, or died within the Netherlands within the 19<sup>th</sup> and 20<sup>th</sup> centuries by matching and harmonising person observations from civil registry records. A combination of several person observations that presumably refer to the same person is called a person reconstruction. The initiators of LINKS chose to base their person reconstructions on the civil registry, as historically willingness to register was high, checks on the registration practice stringent, and double storage meant that the source survived bombardments, fires, floods, and other mishaps (Vulsma, 1988).

Although the civil registry contains a rather complete recollection of all vital events within the borders of the Netherlands, it can be tricky to identify to which person each event belongs. The civil registry chronologically lists all births, marriages, and deaths that occurred in a municipality. Persons are only followed passively, and are

---

<sup>1</sup> See <https://WieWasWie.nl>

<sup>2</sup> ‘Linked Data’, W3CWiki (24 September 2023), <https://www.w3.org/wiki/LinkedData>.

<sup>3</sup> We would like to thank Pieter Woltjer for his insightful feedback and comments.

<sup>4</sup> See <https://www.clariah.nl>.

not observed outside of events, so while we get names of index persons and relatives, we cannot know whether and when a previous or next observation occurs (Alter et al., 2009; Gill, 1997; Van den Berg et al., 2021). As a result, any person reconstruction based on person observations in civil records is always to some extent open to interpretation, as we cannot ascertain that all events associated with a person have been added to a reconstruction. Hence provenance data and a transparent pipeline are key.

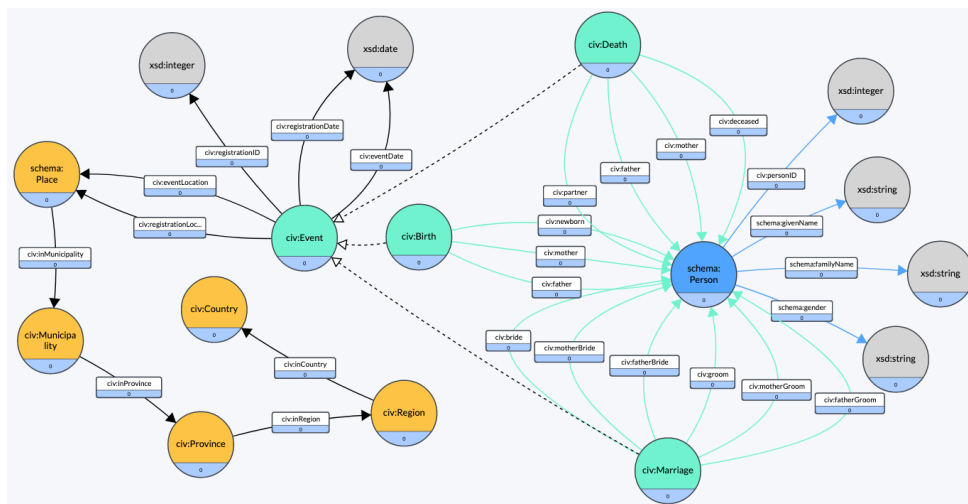


Figure 1: Civil Registries schema for burgerLinker (Raad et al., 2020) (CC0)

Data models help to create reusable software and introduce standards on data provenance. However, the current pipeline (Van Herck and Mourits, 2023) was developed before the creation of PiCo. The Civil Registries schema in figure 1 was purpose-built resulting from ad hoc needs for matching person observations with the custom-made tool burgerLinker.<sup>5</sup> burgerLinker was designed to solve computational problems in comparing millions of name combinations. burgerLinker generates person reconstructions regardless of minor differences in name spelling using the Levenshtein distance to measure similarity between names on two different records (Raad et al., 2020). In order to retrieve unique matches, multiple names on civil certificates are matched, for example ego, father, and mother. The matching strategy in figure 2 demonstrates how person names can be compared for seven types of links, six of which are included in burgerLinker.<sup>6</sup>

Because person observations are converted and checked before person reconstructions can be generated, this schema increases the overhead. In addition, burgerLinker cannot be generalised. Any variation in other sources or similar sources from other countries will likely cause errors. Moreover, comparing output with other matching algorithms is cumbersome, as each package describes provenance slightly differently in their metadata.

If we were to adopt the PiCo standard, dates will be standardised to YMD ISO 8601, names are edited according to the rules of the Person Name Vocabulary (PNV) and Schema.org, certificate types are listed in PiCo terminology, relations are listed in Schema.org, and roles according to PiCo terminology (Woltjer et al., 2024). However,

<sup>5</sup> Joe Raad, 'CLARIAH/burgerLinker', Java (2021; repr., CLARIAH, 30 November 2023), <https://github.com/CLARIAH/burgerLinker>

<sup>6</sup> Within D-M is currently not included.

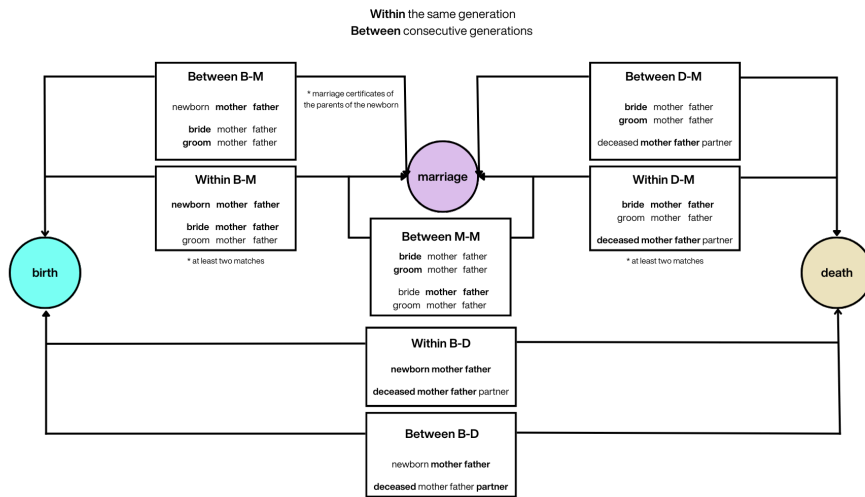


Figure 2: burgerLinker matching strategy (Van Herck and Mourits, 2023). Image reproduced with permission from the copyright owner.

other parties involved in historical person reconstruction have generally used their own schemas and developed their own algorithms to match historical person records, which has hampered the interoperability and reusability of software (Mourits et al., 2023).

The PiCo vocabulary will supersede the Civil Registries vocabulary specifically created for burgerLinker. When creating burgerLinker, no existing ontology fully covered what we needed, therefore we created the Civil Registries vocabulary. The Civil Registries vocabulary was kept simple, because creating a proper schema would be another project entirely. burgerLinker’s architecture was designed to incorporate an ontology with as little effort as possible. The adoption of the ‘new’ PiCo vocabulary, will therefore come with minimal adaption cost as the community has anticipated the arrival of a more fully fletched ontology to describe person observations.

We note three conceptual differences between PiCo and the Civil Registries vocabulary that are important to implement anew in burgerLinker. First, unlike the Civil Registries vocabulary, PiCo discerns the date of the registration of the event and the actual event. While a marriage is usually registered on the same day, there can be a time difference in registration of death and birth. This seemingly small distinction in time is of crucial importance when registrations of baptisms are used to estimate birth dates and in research focusing on infant mortality. Second, PiCo’s elaborate modelling of family relationships and roles supersedes the Civil Registries schema. Third, while the Civil Registries vocabulary was designed with vital event registers in mind, PiCo was designed to be compatible with a wide variety of historical documents, making PiCo more versatile and interoperable.

To illustrate some of the more detailed differences between the PiCo and Civil Registries vocabularies, we created Figure 3, showing the properties relating an event to a person; picot:newborn refers to the role of a newborn on a birth certificate, while picom:deceased has become a property of a person observation, schema:spouse refers to both bride and groom, and schema:parent replaces mother and father. The Civil Registries schema implies gender within the language used to describe the role, whereas

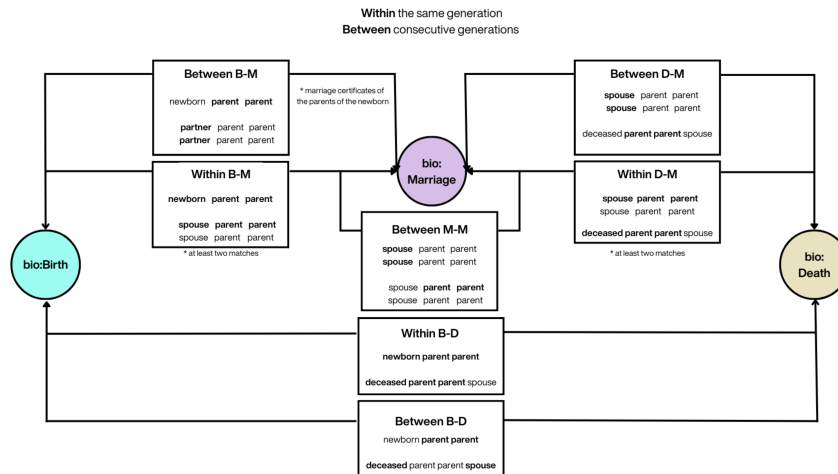


Figure 3: burgerLinker matching strategy remodeled according to PiCo

information on a person's gender in PiCo is explicated via the property schema:gender.

To illustrate the difference in handling names between the Civil Registries Schema and PiCo, figure 4 provides an example of a marriage certificate.<sup>7</sup> While the Civil Registries schema uses the Schema ontology for names, the PiCo model extends Schema with the Person Name Vocabulary (PNV) to include peculiarities of Dutch names.<sup>8</sup> Similarly PiCo allows for other regional and language specific additions to accommodate the expression of names. burgerLinker already dropped the prefix of Dutch surnames for matching, as prefixes are generally seen as separate from the surnames in the source material and can introduce spelling variations in the full sdo:familyName. This split between prefixes and surnames becomes easier when PiCo is implemented. Specifically, burgerLinker can use PiCo's distinction between pnv:surnamePrefix and pnv:baseSurname for family names. These are all minor tweaks to burgerLinker, but would make the program much easier and less labour-intensive to use.

Provenance is central to the PiCo data model, whereas the Civil Registries schema only indirectly links to the source through a civ:registrationID. We created figure 5 to illustrate the PiCo model and to show the link between person observations and archive components in PiCo.<sup>9</sup> This figure illustrates the relation between a node representing the PersonObservation on the left (in blue), with all its potential properties. Moreover, PersonObservation is linked to an ArchiveComponent (in yellow) on the right side of the figure with a property from the PROV-O ontology, specifically designed to capture provenance. This relation is called the 'hadPrimarySource' property. One of the properties describing the ArchiveComponent is the URL that points to the web

<sup>7</sup> CBG, 'PiCo/examples/various-sources/huwelijksakte.ttl', 14 December 2024 <https://github.com/CBG-Centrum-voor-familiegeschiedenis/PiCo/blob/a01283ffc1469285c2dcb280e32842b616187791/examples/various-sources/huwelijksakte.ttl#L77-L88>; CBG, 'PiCo', 20 september 2024, <https://github.com/CBG-Centrum-voor-familiegeschiedenis/PiCo/blob/main/PiCo%20Specificatie.md>.

<sup>8</sup> Schema.org, 9 January 2024, <https://schema.org/>. Lodewijk Petram, Elvin Dechesne, and Gijsbert Kruithof, 'Person Name Vocabulary', Person Name Vocabulary, 1 July 2019, <https://www.lodewijkpetram.nl/vocab/pnv/doc/>.

<sup>9</sup> Visual created in Grafo (<https://grafo>).



to relate the established PersonReconstruction to the PersonObservations. Because of all these links, a PersonReconstruction can be traced back to its original source, constituting a provenance trail containing all the steps from the research result to the sources it was based upon. In this way we are describing the knowledge production process, which enables research replication in the future. (Stapel and Zandhuis, 2024)

### 3 Identifying Person Reconstructions

The second challenge is generating a unique identifier for every person. When civil registries were introduced in the late nineteenth century, governments did not assign identifiers. The civil registry never followed persons over time and 'only' reports on the births, marriage, and deaths that occurred that year within a municipality. However, when the civil registry for a province or the entire country is combined, the person observations can be used to reconstruct life courses and families. In order to keep track of these person reconstructions, we suggest to create a unique identifier for each person reconstruction linking to all observations of that individual.<sup>10</sup> Such a identifier helps to keep track of provenance on changes in sets of observations that 'reconstruct' a person's life course, including edits or dissolution of such a reconstruction.

Generating person identifiers on a national scale over two centuries is no small feat, especially across borders and sources. It requires a balancing act between a stringent protocol stable between different versions of the data and a system tailored to continuous changes. To compare person reconstructions between versions, identifiers cannot be generated randomly. This also means that the name of a person reconstruction cannot rely on the combination of matched certificates, as this would not allow reconstructions to be checked for merging and splitting. PiCo enables us to model person reconstructions with a unique identifier and provenance to the underlying observations.

However, tracking all mutations would generate an unfathomable change log. We suggest using an authority list because some historical person records are more useful than others to determine which unique persons lived in the past. A person reconstruction would receive an identifier based on the highest authority certificate. For the civil registry a logical authority list would be first birth, then death, and finally marriage certificates. This system allows for person identifiers that can be traced whilst allowing for changes between versions. For example, if a person reconstruction contains two sisters named Anna and Anna Maria, it should be split in two separate person reconstructions. One reconstruction with the identifier of the birth certificate of Anna to identify the first person reconstruction and another with an identifier from the remaining certificate with the highest authority to identify Anna Maria. An authority list can also be a set of thoroughly checked person reconstructions. In that case, an alternative to the previous logical historical identifier would be to assign serial identifiers. When mistakes in links are discovered the serial identifier would need to be changed and could possibly result in multiple identifiers for person reconstructions with updated life courses.

Historical person identifiers can be maintained at any level, but ideally at the national level. In the Netherlands, the Center for Family History (CBG) is aiming to become the central network partner that identifies historical persons and assigns them historical person identifiers. In Belgium a logical approach to identifiers would be to transpose the current logic of national person identification numbers (*rijksregisternummers*) to the

---

<sup>10</sup> Al Idrissou, 'C2RC', Python, 26 December 2023, <https://github.com/CLARIAH/C2RC>

past, whereby the date of birth (YY.MM.DD) is followed by a serial number of three digits according to the order of registration of a person with even numbers for female persons and uneven numbers for male persons. The final two digits are calculated by dividing the birth date and serial number by 97 as a checksum. Persons born in or after the year 2000 are assigned an additional digit (2) at the start of their national person identification number. With each century the order of registration begins again, which causes problem for collections of historical person data that span multiple centuries. Therefore, the identifier should be expanded to YYYY.MM.DD-XXX.XX.<sup>11</sup>

## 4 Evaluating and Validating Person Reconstructions (C2RC)

Thirdly, we investigate the role of PiCo in the evaluation and validation of the - possibly automatically generated - reconstructions. The Civil Registry Reconstitutions Cleaner (C2RC), developed in the context of CLARIAH similar to burgerLinker, was designed to systematically evaluate record linkage, such as output from burgerLinker. Reconstructing life courses from person observations is a challenge. Information on records is often incomplete or slightly inaccurate. Moreover, explicitly stating why some matches are considered sound, while others are not, supports reproducibility. Therefore C2RC provides an adaptable set of hard rules and soft rules. Hard rules are matches deemed impossible, such as a reconstructed life course containing multiple birth records. Soft rules are more domain-specific and for example specify a person's oldest feasible age or the maximum number of children a person might have. In addition to evaluation, C2RC allows users to improve upon incongruencies in the linkage based on the hard and soft rules. In line with the concept of provenance trails, C2RC monitors edits based on these violations. To track these improvements C2RC uses the VOID+ vocabulary (Idrissou et al., 2022) as shown in Figure 6.<sup>12</sup>

The VoID+ vocabulary documents *the Creation, Manipulation and Evaluation of Links for Reuse and Reproducibility* and indicates which sources and entities are covered by the dataset, the algorithms and criteria used to generate the data, whether links are validated or not, and if for instance person observations are clustered (Idrissou et al., 2022, p. i). The VoID+ vocabulary reuses the existing vocabularies VoID and PROV-O to create provenance trails (Idrissou et al., 2022, p. ii). Figure 6 illustrates the example from the previous section of a person reconstruction that contains the two siblings Anna and Anna Maria. The initial person reconstruction is now identified as the original cluster that is split into two subclusters. For each of these new subclusters VoID+ tracks what cluster this new person reconstruction was derived from and which person observations it is composed of.

VoID+ allows us to keep track of how data is generated, which validations were applied and how person reconstructions can be split or merged. However, this does not resolve issues in identifying person reconstructions between different matching efforts, as validations and changes to the dataset are lost with each iteration. Furthermore, we could end up with an unfathomable change log that makes the dataset harder to work with despite increasing the chances of proper reproducibility. While PiCo does not implement VoID+ as a way to establish provenance trails, it enables including the PROV vocabulary to model the prov:Activity and thus keep track of the research

---

<sup>11</sup> See [https://www.ibz.rrn.fgov.be/fileadmin/user\\_upload/nl/rr/toegang/bestand-rr.pdf](https://www.ibz.rrn.fgov.be/fileadmin/user_upload/nl/rr/toegang/bestand-rr.pdf)

<sup>12</sup> 'VoIDPlus', 1 August 2022.  
<https://github.com/VoIDPlus-owl/EKAW2022/blob/main/voidPlus.owl>.

```

@prefix      rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix      rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix      dcterms: <http://purl.org/dc/terms/> .
@prefix      voidPlus: <http://lenticularlens.org/voidPlus/> .
@prefix      resource: <https://iisg.amsterdam/id/civ/resource/> .
@prefix      dio: <https://w3id.org/dio#> .
@prefix      civ: <https://iisg.amsterdam/id/civ/> .
@prefix      ind: <https://iisg.amsterdam/links/person/> .

resource:Zeeland-Manual-Validation-H3b094c2f2c6e192
{{{}}} {

    ### 2023-07-11 14:49:13.218922
    ind:i-26785
    a
        voidPlus:ResourceCluster,
        voidPlus:SplitCluster;
        ind:i-1_Hf897efcefc13cf5 ;
        ind:i-2_H988ebffe3f7467e .

        voidPlus:hasValidatedSubCluster
        voidPlus:hasValidatedSubCluster

    ind:i-1_Hf897efcefc13cf5
    a
        voidPlus:ResourceCluster,
        voidPlus:SplitCluster;
        ind:i-26785 ;
        ind:p-411799, ind:p-2867421 .

        voidPlus:derivedFrom
        voidPlus:isComposedOf

    ind:i-2_H988ebffe3f7467e
    a
        voidPlus:ResourceCluster,
        voidPlus:SplitCluster;
        ind:i-26785 ;
        ind:p-2526610 .

        voidPlus:derivedFrom
        voidPlus:isComposedOf

```

Figure 6: C2RC export of the manual validation of a person reconstruction. Screenshot from a user session in C2RC.

process of generating a Person Reconstruction as we have illustrated in figure 7.

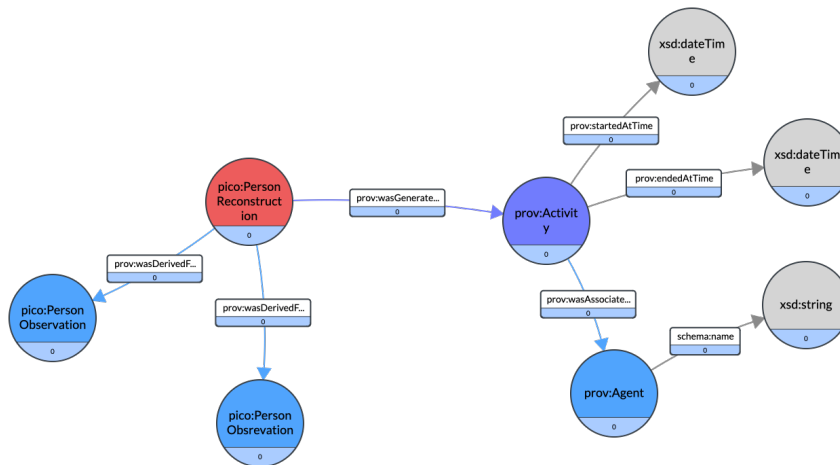


Figure 7: PiCo Person Reconstruction.

As PiCo is expressed using RDF, an alternative and better approach would be to use the standardised Shapes Constraint Language (SHACL) for validation purposes. A shape graph consists of triples describing what patterns can be expected, for example, that a birth year should be expressed as a date in YYYY.MM.DD and the date must fall within a certain time range. This shape graph is compared to a data graph to show incongruencies, similar to the ‘hard’ and ‘soft’ rules in C2RC. SHACL will provide an important contribution as it can be used to see whether data was put in the right data format, and reused in similar tools across the ecosystem. We have included an example of a translation from C2RC’s rule to SHACL in Figure 8. Ideally, there

should be different SHACL shapes for ‘hard’ rules and ‘soft’ rules, standardising current evaluation purposes. Internationally providing these SHACL rules holds huge potential as most evaluation is now done within countries on country-specific linkage methods.

<pre> :person_reconstruction a sh:NodeShape ;   sh:targetClass pico:PersonReconstruction ;   sh:property [     sh:path prov:wasDerivedFrom ;     sh:maxCount 1 ;     sh:node :birth_observation ;   ] . </pre>	<p>“A Person Reconstruction was derived from at most 1 specific Person Observation.”</p>
<pre> :birth_observation a sh:NodeShape ;   sh:targetClass pico:PersonObservation ;   sh:property [     sh:path prov:wasDerivedFrom ;     sh:maxCount 1 ;     sh:node :birth_certificate ;   ] . </pre>	<p>“This specific Person Observation was derived from at most 1 specific Archive Component.”</p>
<pre> :birth_certificate a sh:NodeShape ;   sh:targetClass schema:ArchiveComponent ;   sh:property [     sh:path schema:additionalType ;     sh:minCount 1 ;     sh:maxCount 1 ;     sh:hasValue picot:geboorteregistratie ;   ] . </pre>	<p>“This specific Archive Component is of type <i>birth certificate</i>.”</p>

Figure 8: SHACL example for PiCo. Image reproduced with permission from the copyright owner.

When the indexed civil registry data is remodelled into PiCo, the Civil Registry Reconstitutions Cleaner (C2RC) needs an extensive update. C2RC relies on the graph structure to evaluate person reconstructions against a set of rules, changes to the graph structure would require changes to C2RC. In addition to the benefits of SHACL for record linkage evaluation, SHACL would also make C2RC interoperable with other software components.

## 5 Conclusion

This paper describes potential benefits of applying a standardised data model to an existing pipeline. First we discussed how PiCo (Woltjer et al., 2024) can impact and expand burgerLinker, the software that generates person reconstructions. The second challenge is identifying person reconstructions by generating a unique and persistent identifier. Finally, we explained how a remodel of the indexed civil registry enables a standardised SHACL implementation of the Civil Registry Reconstitutions Cleaner (C2RC), a tool to evaluate and validate person reconstructions. We conclude that PiCo makes each component of the pipeline reusable by improving the interoperability of datasets and algorithms for historic archival sources containing person observations.

The model is only the beginning of expanding across time, space, and source types. PiCo as a shared data standard allows us to further integrate tools into an ecosystem. Anyone can select only part of the ecosystem for their project as long as they adhere to the standard. The tools that are already developed based on the standard can be further integrated into other pipelines. An additional advantage in saving both input and output, is that by using provenance trails, provenance is saved independent from

tooling (Stapel and Zandhuis, 2024).

The updated pipeline allows us to publish the most extensive version of the linked civil registry to date. Implementing major changes for both burgerLinker and C2RC depends on future project funding. Ideally, we make all tools schema agnostic, or at least implement the PiCo model and translate C2RC's evaluation rules to SHACL. Schema agnostic tools can be used for slightly different sources, or for different schemas adapted to similar sources in other countries. The LINKS pipeline could serve as a proof of concept and practical implementation to create person reconstructions for the civil registry. In the future we want to use a more diverse set of sources. By incorporating PiCo, Openarchieven.nl shows how archival data can be provided openly and research ready, building an important bridge between heritage and research. New projects could take the development a step further to link and analyse large amounts of data such as historical sources, texts, and images.<sup>13</sup>

For the Netherlands, the population registers ('bevolkingsregisters') and personal cards ('persoonskaarten') first come to mind, but more elaborate sources like notarial deeds could be included. Current projects about causes of death (Janssens, 2021; Janssens and Devos, 2022), and inheritance (de Vicq and Peeters, 2020) and income taxes<sup>14</sup> can make use of PiCo to enable linkage and result in analyses of a larger combination of variables. People migrate and relate to people in other countries, so data does not stop at the Dutch border. PiCo has chosen international standards such as PROV-O to facilitate linking persons across countries.<sup>15</sup> In the Dutch colonial context, a reconstruction of life courses for Suriname and the Caribbean (HDSC) has been undertaken (Raaijmakers, 2024; Van Galen et al., 2023). We are not aware of person reconstructions in Belgian or Luxembourgish projects.

While indexed civil registries for Belgium have only been made available online from 2017 onwards, 30 years of transcription precede the current LINKS dataset(s).<sup>16</sup> Unlike wiewaswie.nl, genealogie.arch.be separates digitised certificates from persons in the database. Both countries have asked volunteers to transcribe and index the civil registry. Scanned Belgian sources containing person observations are available via Family Search, while person observations have sporadically been transcribed manually by volunteers resulting in substantial differences in availability by region and research project.<sup>17</sup> Furthermore, person observations can be searched via the National Archive website and are not available as a complete dataset for researchers.<sup>18</sup> One Belgian database similar to LINKS worth mentioning is the Antwerp COR-Database built by the Leuven Research Group Family and Population Studies (FAPOS) containing "longitudinal and intergenerational data at the individual level".<sup>19</sup>

AI can ease the workload for data entry from scanned sources with Handwritten Text Recognition (HTR) or Optical Character Recognition (OCR), and Named Entity

---

<sup>13</sup> Thijs van der Veen, 'SSHOC-NL infrastructuur ontvangt 15,2 miljoen', *Huygens Instituut (blog)*, 20 February 2023, <https://www.huygens.knaw.nl/sshoc-nl-infrastructuur-ontvangt-152-miljoen/>.

<sup>14</sup> 'HIP-NL: A Historical Income Panel for the Netherlands', PDI-SSH, 2021. <https://pdi-ssh.nl/nl/2021/11/gehonoreerde-projecten-2021-call/>.

<sup>15</sup> 'The PROV Namespace', The PROV Namespace, 19 May 2013, <https://www.w3.org/ns/prov>.

<sup>16</sup> Mourits, Rick J., Kees Mandemakers, Fons Laan, and Richard L. Zijdeman. 'Cleaned Civil Registry, Netherlands'. IISH Data Collection, 12 July 2023. <https://datasets.iisg.amsterdam/dataset.xhtml?persistentId=hdl:10622/0N0SRV>.

<sup>17</sup> An overview of finished and ongoing projects is available here: <https://search.arch.be/en/zoeken-naar-personen/projecten>

<sup>18</sup> See <https://genealogie.arch.be/>.

<sup>19</sup> See <https://ehps-net.eu/databases/antwerp-cor-database>.

Recognition (NER) to extract person names and model them in line with the PiCo ontology. In turn, the PiCo data model eases the workload by ensuring that burgerLinker can be reused in the Belgian context, while also ensuring that new techniques to create person reconstructions developed in Belgium can be reused in the Netherlands.

## References

- George Alter, Isabelle Devos, and Alison Kvetko. Completing life histories with imputed exit dates: A method for historical data from passive registration systems. *Population*, 64(2):293–318, 2009. URL <https://www.jstor.org/stable/27736142>.
- A. de Vicq and R. Peeters. Introduction to the Tafel v-bis dataset: Death duty summary information for the Netherlands, 1921. *Research Data Journal for the Humanities and Social Sciences*, 5(1):1–19, September 2020. doi: 10.1163/24523666-bja10007. URL [https://brill.com/view/journals/rdj/5/1/article-p1\\_1.xml](https://brill.com/view/journals/rdj/5/1/article-p1_1.xml).
- Antoinette Fauve-Chamoux, Ioan Bolovan, and Sølvi Sogner. *A global history of historical demography*. Peter Lang, 2016. ISBN 978-3-0343-2303-1 978-3-0352-0331-8 978-3-0343-1420-6 978-3-0343-2304-8. URL <https://www.peterlang.com/document/1053040>.
- Bram Gaakeer, Remco de Boer, Enno Meijers, Sjors de Valk, Marcel Ras, Tamara van Zwol, Joost van der Nat, Annelot Vijn, Willem Melder, Erik van den Bergh, Steven Ham, Frans van der Zande, Laurents Sesink, Marcus Cohen, Ivo Dahlmans, Michelle Boon, Gijs Broos, and Mark van Hoorn. Digitaal Erfgoed Referentie Architectuur (DERA) - versie 4.0, October 2021. URL <https://doi.org/10.5281/zenodo.5562062>.
- Richard D Gill. Nonparametric estimation under censoring and passive registration. *Statistica Neerlandica*, 51(1):35–54, 1997. doi: 10.1111/1467-9574.00036.
- Al Idrissou, Veruska Zamborlini, and Tobias Kuhn. Documenting the creation, manipulation and evaluation of links for reuse and reproducibility. In Oscar Corcho, Laura Hollink, Oliver Kutz, Nicolas Troquard, and Fajar J. Ekaputra, editors, *Knowledge Engineering and Knowledge Management. EKAW 2022, Lecture Notes in Computer Science*, pages 81–96. Springer, 2022. ISBN 978-3-031-17105-5. doi: 10.1007/978-3-031-17105-5\_6.
- Angélique Janssens. Constructing ship and an international historical coding system for causes of death. *Historical Life Course Studies*, 10:64–70, 2021.
- Angélique Janssens and Isabelle Devos. The limits and possibilities of cause of death categorisation for understanding late nineteenth century mortality. *Social History of Medicine*, 35(4):1053–1063, 2022.
- Kees Mandemakers, Jan Hornix, Rick J Mourits, Sanne Muurling, Corinne Boter, Ingrid K Van Dijk, Ineke Maas, Bart Van de Putte, Richard L Zijdeman, Paul Lambert, Marco H D Van Leeuwen, Frans W A Van Poppel, and Andrew Miles. *HSNDB Occupations*. IISH Data Collection, 2020. doi: 10622/88ZXD8. URL <https://hdl.handle.net/10622/88ZXD8>.
- Kees Mandemakers, Gerrit Bloothoof, Fons Laan, Joe Raad, Rick J. Mourits, and Richard L. Zijdeman. A system for historical family reconstruction in the Netherlands. *Historical Life Course Studies*, 13:148–185, 2023. doi: 10.51964/hlcs14685.

- Albert Meroño Peñuela, Victor De Boer, Marieke Van Erp, Richard Zijdeman, Rick Mourits, Willem Melder, Auke Rijpma, and Ruben Schalk. CLARIAH: Enabling interoperability between humanities disciplines with ontologies. In *Applications and practices in ontology design, extraction, and reasoning*, pages 73–90. IOS Press, 2020. doi: 10.3233/SSW200036. URL <https://ebooks.iospress.nl/doi/10.3233/SSW200036>.
- Rick J Mourits, Tim Riswick, and Rombert Stapel. Common language for accessibility, interoperability, and reusability in historical demography. In B Steffen, editor, *International Conference on Bridging the Gap between AI and Reality. AISoLA 2023*, Lecture Notes in Computer Science, pages 10–29. Springer, 2023. doi: 10.1007/978-3-031-73741-1\_2.
- Rick J Mourits, K. Mandemakers, F. Laan, C. Munnik, and K. Meijer. *HSNDB Standardisation Tables*. IISH Data Collection, 2024. doi: 10622/IKB8HO. URL <https://hdl.handle.net/10622/IKB8HO>.
- NDE and ministerie van OC&W. Nationale strategie digitaal erfgoed, November 2024. URL <https://doi.org/10.5281/zenodo.14237069>.
- Joe Raad, Rick Mourits, Auke Rijpma, Ruben Schalk, Richard Zijdeman, Kees Mandemakers, and Albert Merono-Penuela. Linking dutch civil certificates. In *Third Workshop on Humanities in the Semantic Web. WHiSe 2020*, pages 47–58. CEUR-WS, 2020.
- Wouter Raaijmakers. A matter of migration? migration patterns of formerly enslaved stations in the post-emancipation caribbean, 1863–1909. *Journal of Caribbean History*, 58(1):28–53, 2024.
- Rombert Stapel and Ivo Zandhuis. Linked Data for modelling and replicating the knowledge production process in data-driven humanities research. *Digital Scholarship in the Humanities*, pages i100–i107, July 2024. ISSN 2055-7671. doi: 10.1093/llc/fqae038. URL <https://doi.org/10.1093/llc/fqae038>.
- Niels Van den Berg, Ingrid K Van Dijk, Rick J Mourits, P Eline Slagboom, Angelique Janssens, and Kees Mandemakers. Families in comparison: An individual-level comparison of life-course and family reconstructions between population and vital event registers. *Population Studies*, 75(1):91–110, 2021. doi: 10.1080/00324728.2020.1718186.
- Coen W Van Galen, Rick J Mourits, MT Rosenbaum-Feldbrügge, M A.B., Jasmijn Janssen, Björn Quanjer, T van Oort, and Jan Kok. Slavery in Suriname. A reconstruction of life courses, 1830–1863. *Historical Life Course Studies*, 13:191–211, July 2023. ISSN 2352-6343. doi: 10.51964/hlcs15619. URL <https://hlcs.nl/article/view/15619>.
- Sytze H. J. Van Herck and Rick J. Mourits. Upcycling the dutch civil registry using linked data. In *Legacy4Reuse Criteria and Methods for Upcycling Data Collections in Social and Economic History*, 2023. URL <https://pure.knaw.nl/portal/nl/activities/upcycling-the-dutch-civil-registry-using-linked-data>.
- Olga Van Oost and Alexander Vander Stichele. Erfgoedonderzoek en kennisontwikkeling. Inspiratienota., June 2024. URL <https://faro.be/publicaties/erfgoedonderzoek-en-kennisontwikkeling-inspiratienota>.

L. Van Toor, J. Claij-Swart, R. Van Gaalen, R.J. Mourits, and R.L. Zijdemán. End report 'Historical Sample of the Netherlands (HSN, task 2.3)' within the ODISSEI Roadmap project. Technical report, HSNDB and CBS, 2022.

Rudolf Ferdinand Vulsmá. *Burgerlijke stand en bevolkingsregister*. Centraal Bureau voor Genealogie, 1988.

P.E. Woltjer, I. Zandhuis, B. Coret, M. Lindeman, J.D. Balkenende, R.L. Zijdemán, and R.J. Mourits. Persons in context: A model to represent observations and reconstructions of historical persons in linked data. *Historical Life Course Studies*, 14(1):105–125, October 2024. doi: 10.51964/hlcs19312.

