IISH Guidelines for preserving research data: a framework for preserving collaborative data collections for future research
About this publication

IISH Guidelines for preserving research data
A framework for preserving collaborative data collections for future research

SURFfoundation
PO Box 2290
NL-3500 GG Utrecht
T + 31 30 234 66 00
F + 31 30 233 29 60

info@surf.nl
www.surf.nl/en

Authors
Afelonne Doek - International Institute of Social History, Amsterdam
Lex Heerma Van Voss - International Institute of Social History, Amsterdam
Karin Hofmeester - International Institute of Social History, Amsterdam
Jan Kok - International Institute of Social History, Amsterdam
Titia Van Der Werf-Davelaar - International Institute of Social History, Amsterdam

Editor
Keith Russell (SURFfoundation)

SURF is the collaborative organisation for higher education institutions and research institutes aimed at breakthrough innovations in ICT (www.surf.nl/en)
This publication is online available through www.surffoundation.nl/en/publications

© Stichting SURF
August 2010
ISBN 9789078887096

This publication is published under the Creative Commons Attribution 3.0 Netherlands Licence.
# Table of Contents

Management summary: practical guidelines ................................................................. 5
Samenvatting: praktische richtlijnen (Translation of the management summary in Dutch) ............................................................... 9

1 Introduction ................................................................................................................ 13
  1.1 Statement of intent ............................................................................................... 13
  1.2 Purpose and scope of this document .................................................................... 13
  1.3 Applicability ........................................................................................................ 13
  1.4 Terminology ........................................................................................................ 13

2 Private, collective and public research data .............................................................. 15

3 Layered research data ................................................................................................ 17
  3.1 Layer 0. Research data collection ........................................................................ 17
  3.2 Layer 1. Source corresponding data set ............................................................... 17
  3.3 Layer 2. Processed data set ................................................................................ 18
  3.4 Layer 3. Data set for analysis ............................................................................. 18

4 Data curation responsibility flows ............................................................................. 19
  4.1 Who .................................................................................................................... 19
  4.2 Why ................................................................................................................... 19
  4.3 When ................................................................................................................. 20
  4.4 What .................................................................................................................. 22
  4.5 How ................................................................................................................... 23

5 Setting rights on collectively created data sets .......................................................... 25

6 Conclusion .................................................................................................................. 27

Appendix 1. DDI Lite template for research data ........................................................... 29
Appendix 2a. Example of XML metadata for a source-corresponding data set .............. 33
Appendix 2b. Example of XML metadata for a release of a processed data set ............... 35
Appendix 2c. Example of XML metadata for a Data set for analysis ................................. 37
Management summary: practical guidelines

Users and producers of research data tend to believe that anything digital can be kept forever because digital data do not take up physical space in the way paper collections do. This is how they perceive digital reality. However, to data archive managers and their system administrators, digital preservation of large data collections not only entails high infrastructural and maintenance costs but also unforeseeable, long-term costs in managing ongoing technical obsolescence.

We see that the data archiving community is overwhelmed by the exponential growth of digital data collection for research. The scientific communication process is opening access to research data. Traditionally, the data published in research reports used to consist of illustrations and examples taken from a much more encompassing and complete data collection, that lay at the basis of the research effort. As data are increasingly collected in digital form, they become much easier to reproduce, redistribute and reuse. The demand for access to the underlying data sets is growing and the ensuing need to validate research data imposes itself.

In a digital form, research data are also easier to exchange, share and manipulate during the research process itself. Increasingly, researchers work together in online research environments to collect, process and interpret data. In doing so, they recognize the need to agree on practices, standards and proper documentation in order to ensure the same quality level of their shared data collection.

We are now seeing that all stakeholders involved in the research data lifecycle are imposing their (sometimes conflicting) conditions and requirements on the data, putting much strain on the overall flow. Funding agencies demand that research data, collected with their funding, are made available to the scientific community. Researchers on the other hand demand exclusive rights to their data for the duration of their research. Publishers in turn demand exclusive availability and access to the data for the peer-review process and for publication on their web portals. Reader preferences for publications that include data sets will influence their citation behaviour, which in turn will trigger authors to provide the data together with their publication. More exposure of the data will lead to demands for improvement of quality standards. The more data that becomes available, the stronger the demands of data archive managers will be for selection criteria, normalized data sets and retention/disposition rules.

Considering the observations made above, we recognize that guidelines for research data should take account of the whole picture even if they address only one specific aspect of the whole. In devising these guidelines for the preservation and availability of collaborative data collections, we have therefore worked from the perspective of the research process as a whole. We have looked at the actors involved in this process and their roles and responsibilities.

Our guidelines highlight the iterative process of data collection, data processing, data analysis and publication of (interim) research results. The iterative process is best analyzed and illustrated by following the dynamics of data collection in online collaboratories. The production of data sets in such large scale data collection projects, typically takes a lot of time, whilst in the meantime research may already be performed on data sub-sets. If this leads to a publication a proper citation is required. Publishers and readers need to know exactly in what stage of the data collection process specific conclusions on these data were drawn. During this iterative process, research data need to be maintained, managed and disseminated in different forms and versions during the successive stages of the work carried out, in order to validate the outcomes and research results. These practices drive the requirements for data archiving and show that data archiving is not a once off data transfer transaction or even a linear process. Therefore from the perspective of the research process, we recommend the interconnection and interfacing between data collection and data archiving, in order to ensure the most effective and loss-less preservation of the research data.

Researchers working together in online collaboratories need organizational capacities and infrastructural facilities that often surpass the capabilities of individual researchers or even
We see natural alliances and cooperation taking place between data collecting projects and data archives that provide infrastructural facilities. Typically, this type of cooperation occurs voluntarily in specific knowledge domains, where repositories and researchers have established long-standing trust relationships. This type of cooperation is a precondition to closely interconnect and interface data collection and data archiving practices.

The selection of data takes place during the data collection stage of the research process and follows the best practices of research methodology. Quality assessment takes place during the peer-review and evaluation stages. These are the mechanisms and instruments by which researchers manage and control their research data. As a rule, researchers will expect data archives to preserve their data sets, without applying additional selection criteria or carrying out data quality assessments. Therefore, from the perspective of the researcher, who is the customer of the data archive in both his role as a data producer and data consumer, we recommend that data archives do not apply selection criteria to the data outside of the research process. This way the research community stays in control and is best served.

Research projects are usually short-lived compared to longer-term data collection efforts, which in turn are short-lived compared to data archives with a mission to provide long-term access to research data. It is therefore necessary for data archives to collect sufficient information about the data set (metadata) to ensure access and reuse of the data in the long term. Typically, these are contextual metadata (about the research project and the methodology) and technical metadata (about the encoding of the data, the processing, the visualization, etc.). Our guidelines stress the importance of best practices in versioning and in ensuring the referential integrity of data sets. The assumption is that, if applied, these best practices will lead to well-behaved data sets and will reduce much of the existing data redundancy. Less copying and duplication of data sets during citation and reuse and more stable linking to the original, authoritative data set in the repository, will improve the research process significantly. These best practices still need to be worked out in more detail, during the ongoing cooperation between researchers and data archives. These practices should serve the long-term purpose of scientific research, and not be inhibited by the short-term constraints of research projects, data collaboratories or archives.

Besides best practices across all disciplines, we also need to look into the nature of discipline specific research data. Close cooperation with historians shows how they digitize data, obtained from paper-based sources, and build source corresponding data sets. Massive amounts of data are extracted from the original sources and digitized, without any prior interpretation. This enables different, consecutive historical interpretations of the same data, without having to go back to the original paper-version. In fact, the ongoing digitization of cultural heritage and scientific collections offers the promise of a data mining paradise for historians. For the historian, digitization and building robust source corresponding data sets are important long-term research investments. Clearly such investments are not made by individual researchers or archives, but require a sustained shared approach by academic and heritage institutions.

Releases of source corresponding data sets (layer 1) and releases of processed data sets (layer 2) on which several research publications are based, which is usually the case with large data collection collaboratories, should be managed by the collaboratory managers. These managers can best decide when the dataset is coherent enough for release, which data can be released and in what form (anonymized, standardized, etc.). The release from data layer 1 and data layer 2 should preferably be deposited in a data archive for long-term preservation. In this way the different parties with an interest and responsibility (such as publishers, authors, etc.) in the data, do not each need to keep and make available a copy of the data set releases. Again well defined organizational responsibilities and good logistics can reduce much redundancy in the storage and distribution of research data.

Data sets for analysis are published together with the corresponding research publications and there is a one-to-one relationship between both. Data sets for analysis are as small as possible, referring mainly to the layer 1 and layer 2 data sets on which they are based. Additional, external data may have been added and should be specified. The metadata describing
the research context and specifying the queries performed or data mining algorithms applied, should accompany this data set. The metadata could best be archived together with the publication, for example in the institutional repository of the researcher or the national deposit library.

Rights on data and data privacy rulings complicate the availability, access and reuse of research data. **In the framework of collaboratories for data collection there are however new opportunities and incentives for reaching agreement on how to handle data rights. We advise to adopt an open data licensing approach.** We also advise to respect an acceptable embargo period during which researchers can collect, process and research the data. Our data layer model provides room for differentiating the access embargo regimes according to the different layers. It could be agreed for example that releases of the source corresponding data layer become open access immediately upon release. This needs further exploration.

Last but not least, **we strongly feel that practical guidelines should be geared towards identifying the requirements, use cases and data flows necessary for interfacing collaboratories and data archives – as this is the only practical way to manage the controlled transfer, preservation and availability of research data.** Please read this document as a preliminary effort in this direction. We hope to continue this effort with the help of many others who share our vision and approach.
Samenvatting: praktische richtlijnen
(Translation of the management summary in Dutch)

Gebruikers en samenstellers van onderzoeksgregevens zijn geneigd te denken dat alles wat digitaal is, voor altijd kan worden bewaard. Digitale gegevens nemen immers geen fysieke ruimte in zoals dat bij gegevens op papier het geval is. Voor hen is dit de digitale werkelijkheid. Beheerders van gegevensarchieven en de systeembeheerders daarvan weten echter dat het digitaal bewaren van grote gegevensverzamelingen niet alleen hoge kosten voor infrastructuur en onderhoud met zich meebrengt maar ook gepaard gaat met onvoorzienbare langetermijnkosten die voortvloeien uit de noodzaak het voortdurende probleem van technische veroudering te ondervangen.

De gemeenschap van gegevensarchivering wordt bedolven onder een exponentiële groei van digitale gegevens voor onderzoek. In het wetenschappelijke communicatieproces worden onderzoeksgregevens vrijer toegankelijk. De gepubliceerde gegevens uit onderzoeksrapporten bestonden traditioneel uit illustraties en voorbeelden die waren ontleend aan een gegevensverzameling die veel omvangrijker en vollediger was en die aan de onderzoeksimpuls ten grondslag lag. Nu gegevens steeds meer in digitaal vorm worden verzameld, zijn ze veel gemakkelijker te reproduceren, te distribueren en te hergebruiken. De vraag naar toegang tot de onderliggende gegevenssets neemt toe en de behoefte aan validatie van onderzoeksgregevens dient zich aan.

Tijdens het onderzoeksproces zelf zijn onderzoeksgregevens in digitale vorm bovendien gemakkelijker uit te wisselen, te delen en te manipuleren. Onderzoekers werken in toenemende mate samen in online onderzoeksomgevingen om gegevens te verzamelen, te verwerken en te interpreteren. Gaandeweg beseffen zij dat het noodzakelijk is overeenstemming te bereiken over methoden, standaarden en het juiste documentatie om voor hun gedeelde gegevensverzameling een consistente kwaliteit te garanderen.

Al degenen die betrokken zijn bij de levensduur van onderzoeksgregevens stellen hun eigen (soms tegenstrijdige) voorwaarden en eisen aan de gegevens en leggen aldus veel druk op de totale gegevensstroom. Financierende instanties eisen dat onderzoeksgregevens die via hun fondsen zijn verzameld voor de wetenschappelijke gemeenschap beschikbaar worden gesteld. Anderzijds eisen onderzoekers exclusieve rechten op hun gegevens voor de duur van hun onderzoek. Op hun beurt eisen dat gegevens beschikbaar en toegankelijk zijn voor collegiale toetsing en voor publicatie op hun internetportalen. De voorkeur van lezers voor publicaties die gegevenssets bevatten, beïnvloedt dit citergedrag, hetgeen auteurs ertoe aanzet de gegevens samen met hun publicatie aan te bieden. Naarmate de gegevens gemakkelijker toegankelijk zijn, ontstaat een grotere behoefte aan verbetering van de kwaliteitsnormen. Hoe meer gegevens beschikbaar komen, hoe sterker de roep onder beheerders van gegevensarchieven om selectiecriteria, genormaliseerde gegevenssets en regels voor het behouden of vernietigen van gegevens.

In het licht van bovenstaande opmerkingen is het duidelijk dat in richtlijnen voor onderzoeksgregevens de richtlijnen van de totale levensduur van onderzoeksgregevens moeten worden gehouden. Zelfs wanneer de richtlijnen zich slechts op één specifieke aspect van het geheel richten, moeten ze bij de beelding van de werkelijkheid worden gehanteerd. Bij het samenstellen van deze richtlijnen voor het behoud en de beschikbaarheid van gegevensverzamelingen hebben we daarom het onderzoeksproces als geheel als uitgangspunt genomen. We hebben daarbij gelet op de betrokkenen bij dit proces en hun rol en verantwoordelijkheden.

In onze richtlijnen ligt de nadruk op het herhalingskarakter van het verzamelen, verwerken en analyseren van gegevens en de publicatie van (tussentijdse) onderzoekresultaten. Het herhalingsproces kan het beste worden geanalyseerd en geïllustreerd door de dynamiek van gegevensverzameling en in online samenwerkingsverbanden te volgen. De productie van gegevenssets in dergelijke grootschalige projecten voor gegevensverzameling neemt doorgaans veel tijd in beslag. Het onderzoek aan de hand van subsets van gegevens kan dan intussen al in volle gang zijn. Als dit tot een publicatie leidt, is goed citeren noodzakelijk. Uitgevers en lezers moeten precies weten welk stadium van het gegevensverzameling en besluiteloos word het gegevensverzameling en besluiteloos te maken voor de publicatie op hun internetportalen. De voorkeur van lezers voor publicaties die gegevenssets bevatten, beïnvloedt dit citergedrag, hetgeen auteurs ertoe aanzet de gegevens samen met hun publicatie aan te bieden. Naarmate de gegevens gemakkelijker toegankelijk zijn, ontstaat een grotere behoefte aan verbetering van de kwaliteitsnormen. Hoe meer gegevens beschikbaar komen, hoe sterker de roep onder beheerders van gegevensarchieven om selectiecriteria, genormaliseerde gegevenssets en regels voor het behouden of vernietigen van gegevens.
verschillende vormen en versies worden bijgehouden, beheerd en verspreid in de opeenvolgende stadia van het uitgevoerde project, om de uitkomsten en onderzoeksresultaten te valideren. Deze methoden maken gegevensarchivering noodzakelijk. Kinderen weten dat gegevensarchivering geen eenmalige transactie van gegevensoverdracht is, zelfs geen lineair proces. **Vanuit het onderzoeksproces bezien, pleiten wij daarom voor onderlinge verbinding en koppeling van gegevensverzameling en gegevensarchivering om tot een zo doeltreffend mogelijk behoud van onderzoekseigenschappen te komen met een minimum aan gegevensverlies.**

Onderzoekers die in online samenwerkingsverbanden samenwerken, moeten beslissen over organisatorisch vermogen en infrastructurele voorzieningen die het potentieel van individuele onderzoekers of zelfs academische faculteiten vaak te boven gaan. **Er ontstaan natuurlijke allianties en samenwerkingsvormen tussen gegevensverzameling en gegevensarchieven.** Deze vorm van samenwerking vindt meestal op vrijwillige basis plaats in domeinen van specifieke kennis waarin tussen kenniscentra en onderzoekers jarenlange vertrouwensrelaties bestaan. Dit type samenwerking is een eerste vereiste om methoden van gegevensverzameling en gegevensarchivering onderling nauw te verbinden en te koppelen.

De gegevens worden geselecteerd tijdens de fase van gegevensverzameling van het onderzoeksproces en de selectie vindt plaats volgens de beste praktijken op het gebied van onderzoeksmethodologie. De kwaliteitsbeoordeling verloopt via collegiale toetsing en evaluatie. Dit zijn de mechanismen en instrumenten waarmee onderzoekers hun onderzoekseigenschappen beheren en sturen. Onderzoekers verwachten dat hun gegevenssets in gegevensarchieven worden bewaard zonder dat aanvullende selectiecriteria worden toegepast of kwaliteitsbeoordelingen worden uitgevoerd. **Vanuit het perspectief van de onderzoeker, die klant is van het gegevensarchief in zowel zijn rol als producent als in die van consument van gegevens, bevelen wij aan dat gegevensarchieven geen selectiecriteria toepassen op gegevens buiten het onderzoeksproces. Zo blijft de controle bij de onderzoeksgemeenschap rusten en is deze gemeenschap het meest gediend.**

Onderzoeksprojecten zijn meestal van korte duur in vergelijking met het langdurige proces van gegevensverzameling, dat op zijn beurt weer van korte duur is in vergelijking met gegevensarchieven die bedoeld zijn om nog heel lang toegang tot onderzoekseigenschappen te verschaffen. Daarom moeten gegevensarchieven voldoende informatie verzamelen over de gegevensset (metagegevens) om toegang tot gegevens en het hergebruik daarvan op de lange termijn mogelijk te maken. Dit zijn doorgaans contextspecifieke metagegevens (over het onderzoeksobject en de methodologie) en technische metagegevens (over codering van de gegevens en over verwerking en visualisering en dergelijke). **In onze richtlijnen wordt het belang benadrukt van de beste praktijken bij de vorming van versienummers en het garanderen van de referentiële integriteit van gegevenssets. Bij toepassing zullen deze beste praktijken waarschijnlijk leiden tot deugdelijke gegevenssets en zal de bestaande gegevensredundantie voor een aanzienlijk deel kunnen worden teruggedrongen.** Als gegevenssets tijdens citeren en hergebruik minder worden gekopieerd en geduplicateerd, en de koppeling naar de originele, gezaghebbende gegevensset in het kenniscentrum stabiel is, zal het onderzoeksproces aanzienlijk verbeteren. Deze beste praktijken moeten nog nader worden uitgewerkt tijdens de voortgaande samenwerking tussen onderzoekers en gegevensarchieven. De praktijken zijn bedoeld voor wetenschappelijk onderzoek op de lange termijn en de toepassing ervan mag niet belemmerd worden door kortetermijnbeperkingen van onderzoeksprojecten, samenwerkingsverbanden of gegevensarchieven.

Behalve naar beste praktijken voor alle disciplines moeten we ook kijken naar de aard van onderzoekseigenschappen die voor bepaalde disciplines specifiek zijn. Uit nauwe samenwerking met historici blijkt hoe zij uit papieren bronnen verkregen gegevens digitaliseren en bronspecifieke gegevenssets samenstellen. Enorme hoeveelheden gegevens worden zonder voorafgaande interpretatie uit de oorspronkelijke bronnen geëxtraheerd en gedigitaliseerd, zodat verschillende historische interpretaties van dezelfde gegevens mogelijk zijn zonder dat de oorspronkelijke papieren versie hoeft te worden geraadpleegd. Dankzij de voortdurende digitalisering van gegevens met betrekking tot cultureel erfgoed en wetenschappelijk verzamelingen, ontstaat er
voor historici die gegevens zoeken een waar paradijs. **Voor de historicus vormen het digitaliseren en opzetten van omvangrijke bronspecifieke gegevenssets belangrijke onderzoeksinvesteringen op de lange termijn.** Dergelijke investeringen worden natuurlijk niet door afzonderlijke onderzoekers of archieven geleverd maar vergen een aanhoudende gemeenschappelijke inspanning van academische instellingen en erfgoedinstanties.

Het vrijgeven van bronspecifieke gegevenssets (laag 1) en verwerkte gegevenssets (laag 2) waarop verschillende onderzoekspublicaties zijn gebaseerd, wat meestal het geval is bij samenwerkingsverbanden die met grote gegevensverzamelingen werken, moet door de beheerders van de samenwerkingsverbanden worden beheerd. Zij weten als geen ander wanneer de gegevensset voldoende samenhang heeft om te kunnen worden vrijgegeven, welke gegevens kunnen worden vrijgegeven en in welke vorm dit moet gebeuren (geanonimiseerd, gestandaardiseerd enz.). **De vrijgave van gegevens uit laag 1 en 2 moet bij voorkeur plaatsvinden naar een archief waarin de gegevens voor de lange termijn worden bewaard.** Zo hoeven de partijen die belang stellen in de gegevens en die voor de gegevens verantwoordelijk zijn (zoals uitgevers, auteurs enz.), niet elk een exemplaar van de uitgebrachte gegevenssets te bewaren en beschikbaar te stellen. Met goed gedefinieerde organisatorische verantwoordelijkheden en een juiste logistiek kan redundantie in de opslag en distributie van onderzoeksgegevens sterk worden verminderd.

Gegevenssets voor analyse worden samen met de bijbehorende onderzoekspublicaties in een 1-op-1-relatie uitgebracht. **Gegevenssets voor analyse zijn zo klein mogelijk en hebben hoofdzakelijk betrekking op de gegevenssets uit laag 1 en laag 2 waarop ze zijn gebaseerd.** Aanvullende externe gegevens die mogelijk zijn toegevoegd, moeten worden gespecificeerd. De metagegevens die het onderzoeksproject beschrijven en die aangeven welke query’s uitgevoerd of welke data mining-algoritmen toegepast zijn, moeten met deze gegevensset worden meegezonden. De metagegevens kunnen het beste samen met de publicatie worden gearchiveerd, bijvoorbeeld in de kennisbank van het instituut van de onderzoeker of in de nationale bibliotheek waar onderzoeksgegevens worden bewaard.

Rechten op gegevens en uitspraken over privacy van gegevens maken de beschikbaarheid en het hergebruik van onderzoeksgegevens en de toegang daartoe gecompliceerder. **Er zijn echter nieuwe kansen en impulsen om in het kader van samenwerkingsverbanden voor gegevensverzameling tot overeenstemming te komen over het omgaan met gegevensrechten.** Wij raden een benadering van open gegevenslicenties aan. Verder pleiten wij voor een aanvaardbare embargoperiode waarin onderzoekers de gegevens kunnen verzamelen, verwerken en bestuderen. Ons gelaagde model biedt ruimte om de regimes voor toegangsembargo’s in overeenstemming met de desbetreffende lagen te differentiëren. Men zou bijvoorbeeld kunnen afspreken dat gegevens die uit de bronspecifieke gegevenslaag worden uitgebracht, meteen na vrijgave voor iedereen vrij toegankelijk zijn. Dit moet nader worden onderzocht.

Wij zijn er, tot slot, van overtuigd dat praktische richtlijnen moeten zijn afgestemd op het in kaart brengen van de vereisten, use cases en gegevensstromen die noodzakelijk zijn om samenwerkingsverbanden en gegevensarchieven te koppelen, omdat dit de enige praktische manier is waarop de gecontroleerde overdracht, het behoud en de beschikbaarheid van onderzoeksgegevens kunnen worden beheerd. Wij verzoeken u dit document te beschouwen als een allereerste aanzet daartoe en hopen deze aanzet een vervolg te kunnen geven met behulp van vele anderen die onze visie en benadering delen.
1 Introduction

These Guidelines have been developed in response to the SURF Call Setting up Data Collections, within the context of the SURFshare programme 2007-2010, WP7 Permanent Access to Research Data. This publication is part of a series of three reports directed at collecting guidelines from several research disciplines.

1.1 Statement of intent

The International Institute of Social History (IISH) is an organization with the strategic objective to facilitate the collection of data for carrying out the long-term research programmes of Global Labour History and Global Economic History. The collection process is often a collaborative research effort and the IISH aims to become a trusted digital repository (TDR) of data collections that are significant to social and economic history.

The IISH Guidelines formulate best practices that address the preservation challenges common to all research data collected in collaboratories. In the preservation policy outlined here, the life cycle of such collaboratories plays an integral part.

1.2 Purpose and scope of this document

The purpose of this document is to provide guidelines and best practices for the preservation of research data collections. The Guidelines:

- provide a framework for the understanding and increased awareness of preservation requirements for research data collected by collaborations;
- provide concepts needed by non-archival entities (individual researchers, research collaborations, publishers) to be effective participants in the preservation process;
- suggest selection criteria for the long-term preservation of research data;
- provide a framework for identifying requirements, use cases and data flows necessary for interfacing collaboratories and trusted digital repositories.

1.3 Applicability

The Guidelines are specifically applicable to:

1. Collaborations that produce data collections that may need to be preserved by trusted digital repositories;
2. Trusted digital repositories with the responsibility of making research data collections available for the longer term;
3. Researchers who may need to access the data collections from such trusted digital repositories.
4. Publishers who invite peer-reviewers to verify or reproduce research based on data sets stored in trusted digital repositories.

1.4 Terminology

A number of central concepts used throughout the Guidelines are defined here.

**Codebook:** metadata documenting the data per field, used in a database.

**Collaboratory Manager:** fulfils the role of manager of the data collection collaboratory and usually also the role of head of the research team collecting research data.
**Data collection collaboratory:** team of distributed researchers who, generally for the purpose of a specific research project, create data sets collectively (or join individually collected data).

**Data set (or data set):** a collection of data, usually presented in tabular form.

**Data set for analysis:** A data set based on (a subset of) a processed data set and used for analysis in a specific research context. The findings based on research carried out with the data set for analysis are published in scientific publications.

**Legal) Electronic deposit (e-depot):** Deposit is a legal or voluntary requirement that authors or publishers submit copies of their publications to a national repository, usually the national library. The requirement is mostly limited to books and journals, but increasingly involves data sets as well.

**Processed data set:** A data set, based on a source-corresponding data set, on which cleansing techniques and methods, such as anonymization, normalization, harmonization, coding, etc. have been performed.

**Release:** The distribution (whether public or restricted) of an initial or major upgraded version of a data set. Releases are numbered versions identifying successive states of a database at different points in time and containing unique temporal instances of the data set.

**Research data collection:** Data collected by researchers for research purposes. In contrast with data which may also be used for scientific research purposes, but which is collected by organizations who register data in order to carry out their (legal) tasks (Civil data, Land registry data, Office of National Statistics, etc.).

**Scientific) Data archive:** University/academic organisations that aim to preserve data and methods produced by scientific research.

**Source-corresponding data set :** A data set of research data, collected from one or several sources (primary or secondary sources), digitized and stored in a database, but not processed or manipulated in any way yet. The raw data corresponds closely to the data extracted from the original source.

**Trusted (digital) repository:** Scientific/cultural organisation with a mission to collect and archive (digital) assets for the long-term (not only data sets but all kinds of historical sources).
2 Private, collective and public research data

In order to situate our Guidelines in the research landscape, the process of data collection has been visualised (Figure 1). Obviously, many researchers, in particular in the Humanities, still collect their data individually. Depending on the grant conditions, they may transfer their data after the completion of their research project to an e-depot, data archive or trusted digital repository. Some of them may publish the data on personal websites. Data that have become public in such a way, may be reused in the context of other, e.g. collaborative, research projects. These Guidelines do not address the transfer and curation of individually collected data. Instead, they focus on the transfer and curation of collaborative data collections (the central panel of Figure 1). Increasingly, researchers decide to pool (parts of) their data, in order to increase scale and scope of their research, to engage in comparative projects, to benefit from commonly developed tools for metadata description and data analysis, etc.

Figure 1: The dynamic process of data collection

Increasingly, such pooling of efforts takes place online, either in official (often externally sponsored) collaboratories or in makeshift constructions of researchers. Collaboratories can be defined as teams of distributed researchers who, generally for the purpose of a specific research project, create data sets collectively (or join individually collected data). In the collaboratories, the participants generally recognize the need for a proper documentation and permanent preservation of the common database, for future (re)use. However, collecting and sharing data in commonly
agreed standard formats often requires organisational capacities and infrastructural facilities that surpass the possibilities of individual researchers or university departments. Given this situation, data collecting researchers and data archives/trusted repositories can choose to cooperate for mutual benefit. Typically, this type of cooperation occurs voluntarily in specific knowledge domains, where repositories and researchers have established long-standing trust relationships. In such relationships, the archiving institutions assist teams of researchers in deciding on standards, versioning and durable preservation of their data and by providing service infrastructures. Researchers, in return, donate or bequest their data sets to the archiving institutions – thereby adding value to the data collection held at the archive. Clearly, this cooperation is a delicate matter of negotiation. Sharing and preserving research data will always be based on the voluntary participation of individual researchers and setting strict guidelines or preconditions will hamper the process.

Thus, trusted repositories will want to develop good working relations with research collaboratories, in order to ensure dialogue with researchers early in the process of data collection. Researchers, even the individual ones, need to be aware of the archiving requirements (data formats, documentation) for the future preservation of data. Researchers, in turn, will want to know, at an early stage of their work, whether their data are eligible for curation.

Collaboratories intending to transfer their data to a data archive can use these guidelines to check whether their (meta) data meet the minimum requirements. Finally, collaboratories hosted by archiving institution, can work from the onset of their project with the guidelines set out below.
3 Layered research data

Central to our guidelines is the notion of layered data, corresponding to the dynamic data collection taking place in collaboratories. The production of such a collaborative data collection typically takes a lot of time, whilst in the meantime research may already be performed on (temporary) data subsets, which will have to be cited properly in case of publications. Publishers and readers need to know exactly in what stage of the data collection process specific conclusions on these data were drawn. In figure 2 the different levels of data are visualised.

3.1 Layer 0. Research data collection

The first layer concerns data collected by researchers for research purposes. This in contrast to data collected by institutions that collect data as part of their mission (e.g. civil records collected by municipalities, land ownership data by cadastral offices). These data are the raw, collected, digitized or not digitized data, originating from primary and secondary sources such as archival records and publications. Metadata on this level consist of source descriptions (archival or bibliographic).

3.2 Layer 1. Source corresponding data set

The next layer consists of the data set with the initial version of the digitized data collection, in which the raw data (still) correspond closely to the original source and have not yet been processed or manipulated in any way.

A release of the source corresponding data set should always be preserved to ensure future access to the original (source corresponding) data values on which the research, the publications and
consecutive releases were based. As data collection is an ongoing process, it might be necessary for longer-term data collection collaboratories to preserve the initial release of this data set and consecutive releases. The frequency of releases of the source-corresponding data set will follow the logic of data collection. Releases need to be preserved by trusted repositories/data archives. Accompanying metadata at this level will include references to the original source material, collaboratory information and codebooks.

3.3 Layer 2. Processed data set

The following layer of data consists of data sets that contains data on which cleansing techniques and methods, such as anonymization, normalization, harmonization, coding, etc. have been performed. As data cleansing is an ongoing process, it might be necessary for longer-term data collection collaboratories to preserve releases of the data set in consecutive cleansing stages. The frequency of (internal) releases of the processed data set will follow the collaboratory logic of data cleansing. Public releases of this data set will tend to follow the logic of research demands (e.g. pertaining to subsets, such as a completed geographical region or time period) and will need to be preserved in trusted repositories/data archives.

3.4 Layer 3. Data set for analysis

The final layer contains data sets that are (a subset of) data from layer 2 (processed data set) and possibly additional data collected outside the context of the collaboratory. A data set for analysis corresponds to the data set used for producing the final research results as they are made public in a publication. The publication refers to the data set for analysis by providing references to releases of the layer 1 and layer 2 data sets, together with a specification of the subset and of the additional data used, including the queries and all documentation necessary to inspect and/or reproduce specific tables and/or figures. The publication, together with metadata and references to the data set for analysis, is submitted to the (legal) depot.

The metadata comprise a description of the properties and contents of each data set. They should comply with formats for documenting data sets. Since the data sets defined in these guidelines are situated in different layers, metadata are also layered and can refer to underlying metadata, thereby avoiding redundant information. For instance, as the original source material is already described in Layer 1, the general reference in Layer 2 to Layer 1 will ensure that the metadata from layer 1 is included.
4 Data curation responsibility flows

All actors in the research data flow – from collection, through processing, analysis and publication to archiving – need to make informed decisions about what to preserve, when, how, etc. Below we answer the main questions which need to be addressed: Who takes the decisions and who has the responsibility? Why are decisions taken, on the basis of which criteria? When, at what stage of the data collection process are steps for curation required? What is to be preserved – what kind of selection criteria do we need? And, finally, how are research data to be preserved?

4.1 Who

Who are the key actors and what are their responsibilities?

1. If a research project involves relatively long term and/or large scale data collection in the form of a data collection collaboratory, the need to make the data collection process explicit becomes more urgent. This stems from the logic of (long) distance cooperation between scholars. **Collaboratory managers** need to make sure that agreements are made on clearing rights and sharing data (see section 5). Collaboratory managers are usually in a good position to enforce the use of taxonomies, codebooks for documenting variables and agreed standards for the metadata formats. Furthermore, the collaboratory managers are responsible for the safe storage of data (including back-up facilities) and for restricted access of the data in Layer 1 (as agreed upon by the members of the research team) so long as it concerns work in progress. Collaboratory members may also want to perform research, producing data sets for analysis and publications, on the basis of their work in progress. The best practice here is that the data collection **research team** reaches an agreement on releasing the processed data in Layer 2. Releases can contain subsets of the data, e.g. only the finished data, only standardized or anonymized data, or only the data for a particular region or period. Thus, a release is usually less bulky than the complete database. As soon as publications are based on a data set release (layer 1 and layer 2), the release should be stored in a trusted repository or data archive. Depending on decisions made by the team, the releases may also be made public. As soon as the data collection research project is finished, the collaboratory manager is responsible for handing over the data sets (layer 1 and layer 2) to a trusted digital repository.

2. **Publishers** who receive data sets for analysis (layer 3) together with the publications, are responsible for 1) restricted and controlled access to the data during the peer-review process and 2) deposit of the publication and the data set at the national deposit library. Usually, in case of a 'work in progress' publication, data sets will only be accessible to the peer reviewers.

3. **Peer reviewers** receive restricted access to the data set for analysis and may only use the data set to review the publications based on the data.

4. **Archivists and managers of trusted repositories** are responsible for the preservation and long term access to the data set releases as well as the accompanying metadata (including codebooks and possible other documentation).

5. **National deposit libraries and repositories** of research publications are responsible for safekeeping the final data sets for analysis, needed to verify of research results, together with the publication.

4.2 Why

Individual researchers or research teams who want to publish with a journal having a Data Availability Policy (e.g. *American Economic Review*) need to provide at least the final ‘data set for analysis’ pertaining to a particular publication. Such a data set will usually refer to a release of the
processed data set used. Ideally, the release itself is stored and can be inspected at a Trusted Digital Repository.

A number of data collection collaboratories are, from the onset, working towards the integration and improved access of archival data. Clearly, preservation plans are an integral part of their endeavour. Examples are the Founders and Survivors database in Tasmania\(^1\) and the Historical Violence Database\(^2\). Other collaboratories however are primarily aiming to get a common (set of) publication(s) published. Often, there are no clear plans for future storage and an archival organization is only contacted after the project has ended, if ever. This implies extra costs and difficulties for the preservation of the data, as (re)constructing metadata is a laborious task. We therefore advise that trusted repositories and research collaboratories start negotiations regarding data preservation at the earliest possible stage, and that funding agencies make this obligatory.

For archivists and repository managers, the decision to preserve research data should be derived from the mission statement of the archiving organisation. Publishers make decisions regarding the storage of research data based on their desire to enhance the peer-review process or to provide additional content to their subscribers and the general public, depending on both their Data Availability Policy and Open Access policy.

### 4.3 When

When should an archive take measures to preserve the data collected in a collaboratory? Given the cost of permanent storage of digital data, a number of selection criteria can guide decision-making. In essence, all criteria stem from the relevance of the data to research. Scientific ‘relevance’ of new data sets can be said to be proven when publications are based on them. Data sets that are not linked to published research, might after some time, be discarded. We propose three potential situations:

1. The data collection collaboratory is hosted by the archiving institution and is still collecting or upgrading research data. Releases of the data set layers 1 and 2 are created by the collaboratory, generally in order to make possible clear references to the data collection in publications. The trusted repository will take steps from the start, to store the releases for future reference and use, in close cooperation with the collaboratory manager. The archive will preserve all data set releases and accompanying metadata, on the condition that, within ten years, publications or new grant proposals are based on them.

2. The collaboratory is hosted by the archiving institution but has ceased its data collection and processing work, for whatever reason (drying up of funds, etc.). The archive will preserve all data sets and accompanying metadata, on the condition that, within ten years, publications or new grant proposals are based on them.

3. The collaboratory is not hosted by an archiving institution but wishes to deposit its data sets, after the lifetime of the collaboratory. Depending on the archive’s mission the data will be considered more or less relevant. In all cases, however, the relevance of the data sets (measured in publications based on the data sets) and the availability of proper metadata (or the prospects of them becoming available) will help determine whether the data should be stored permanently.

The decision flows related to preserving data from collaboratories are visualized in figure 3.

\(^1\) [www.foundersandsurvivors.org](http://www.foundersandsurvivors.org)

\(^2\) [http://cjrc.osu.edu/index.html](http://cjrc.osu.edu/index.html)
Figure 3: Decision flow on preserving research data

Ingest – decision flow for submitting research data collected in collaboratories

- Release by collaboratory manager to archiving institution
  - Collaboratory has agreement with archiving institution? (yes/no)
    - Yes: Agreement collaboratory with archiving institution
    - No: Data relevant to the archiving institution's mission? (yes/no)
      - Yes: Release layer? (yes/no)
        - Yes: Related release layer preserved? (yes/no)
          - Yes: Proper metadata? (yes/no)
            - Yes: Archiving institution preserves release in TDR with accompanying metadata
              - End
          - No: Feedback to collaboratory manager
        - No: Proper metadata? (yes/no)
          - Yes: Archiving institution preserves release in TDR with accompanying metadata
            - End
      - No: Feedback to collaboratory manager
  - No: Archiving institution preserves release in TDR with accompanying metadata
    - End

Preservation – decision flow for preservation of research data collected in collaboratories

- 10 years after release
  - Publications or new grant proposals are based on the dataset? (yes/no)
    - Yes: Removal of release from TDR by archiving institution
    - No: Permanent preservation of release by archiving institution
      - End
4.4 What

If the decision to preserve the data is positive, then what actually needs to be preserved?

**Preserving the source correspondent and processed data sets (Layers 1 and 2)**
Data archives or trusted repositories will preserve the releases of the main data sets (layer 1 and 2), together with the codebooks, further documentation and the metadata.

**Contextual information:** Metadata on the data collection project or collaboratory provides provenance for the data. It should contain descriptions of why the data was collected, e.g. (official) information about the project/collaboratory aims and objectives and by whom, e.g. information on the research team.

**Documenting data collection (layer 1):** A collection profile should be provided, containing references on the source materials, how the data were collected.

**Documenting data manipulation (layer 2):** To understand what has been done by the collaboratory to process the source corresponding data, a detailed description of methods and techniques applied (e.g. data manipulation like normalization or anonymization) needs to be part of the metadata. The data set release should be documented: providing title, release number, location, date and nature of the release (layer 1 or 2), the data files and the variables used in the research data. A statement must be made on the access rights of the release. Is it available to the general public, or is it for research purposes only? If the release is not public, something should be said about the duration of the restriction, i.e. when the restriction will end. If confidentiality and consent agreements are applicable in the use of research data, this should also be mentioned here.

**Technical documentation of the data set (both layers):** Information on the data files should contain descriptions of the format, size and structure of the data fields, e.g. the technical specifications of the release (number of records, original format of the data, application used, structure and properties of data fields).

**Information on the variables is often provided in a codebook.** A code book ensures that the data can be understood in the long term. It should be explanatory on the data fields and structure, describing the names, labels and descriptions for data elements and any rules relating to the values that are in the data set (a list of codes used in the research, a coding or classification scheme etc.).

**Preserving a data set for analysis (layer 3)**
The data set for analysis is a final processed (subset) of a release on which specific tables or figures in a publication are based. Thus, the metadata should provide the queries that have been applied to transform the release into this data set, as well as the queries underlying the tables and figures included in the publication.

Metadata on the data set for analysis should provide a reference to the release of the underlying data set (layer 2). It should contain a title of the data set for analysis, location, information on the researchers contributing data to this data set for analysis and the research context in which the data set for analysis has been made. A statement must be made on access rights of the data set for analysis, depending on whether it is intended for public or closed release. Is it available to the general public or not? If the data set for analysis is not public, something should be said on the duration of the restriction, i.e. when the restriction will end. If confidentiality and consent agreements are applicable in the use of this data set for analysis, for example if peer reviewers should only use the data set for analysis for reviewing the publication, this should also be mentioned.
Metadata on the data file should contain descriptions of the format, size and structure of the data fields, e.g. the technical specifications of the data set for analysis (number of records, which tables it contains, performed queries and views).

Metadata on the variables should be explanatory on the data fields and structure, describing the names, labels and descriptions for data elements and any rules relating to the values that are in the data set (a list of codes used in the research, a coding or classification scheme etc.). If the variables used in the data set for analysis are no different than the ones used in the release of the underlying data set, a reference to the release is sufficient.

4.5 How

We recommend online access to public releases and data sets for analysis. Rights on the data, licensing issues and other restrictions will vary for each data set but should be made clear to the user. This will also be discussed in section 5 on rights on collectively created data sets. If releases and data sets for analysis are not publicly accessible, restricted access should be provided online for authorized researchers and users.

The data sets and releases should be stored in the open standard XML to enable long term access to the data. It is advisable that national deposit libraries and repositories of enhanced publications do the same with the data sets for analysis. Data stored in XML can be imported in existing software applications if the need for reconstructing the data set in a database environment exists. Storing the data in original formats and software applications is not recommended. This will intensify the storage requirements and procedures without ensuring future access and usability of the data stored. For research data the preservation of the original software and hardware, look and feel and functionality is not a basic requirement. A basic requirement is to be able to reproduce data sets for analysis. Migrating the data sets for future access will be sufficient.

The documentation and metadata that are to be kept with a release or data set for analysis should also be stored in structured XML, preferably in a discipline-specific metadata standard.

The Data Documentation Initiative (DDI) is specifically designed for describing and sharing data within the social sciences. The DDI is an effort to establish an international criterion and methodology for the content, presentation, transport, and preservation of metadata about data sets. For our purposes, we advocate using the DDI-Lite version, which is not a separate specification, but rather a subset of the full DDI model. The selected elements have shown to be either common across standards or critical to best practice in the documentation of social science data.

The five main sections of the Document Type Definition (DTD) for social science data documentation developed in DDI are listed below. These are the highest-level components of any document that will be marked up in compliance with this DTD.

Document Description. Items describing the marked-up document itself as well as its source documents (citation, title, etc.). Element -- optional, not repeatable.

Study Description. Items describing the overall data collection (title, citation, methodology, study scope, data access, etc.). Element -- required, repeatable.

Data Files Description. Items relating to the format, size, and structure of the data files. Element -- optional, repeatable.

---

3 The DDI Tag Library can be found at [www.ddialliance.org/dtd/](http://www.ddialliance.org/dtd/)

4 [http://www.icpsr.umich.edu/DDI/](http://www.icpsr.umich.edu/DDI/)
**Variables Description.** Items relating to variables in the data collection. Element -- optional, repeatable.

**Other Study-Related Materials.** Other study-related material not included in the other sections (bibliography, separate questionnaire file, etc.). Element -- optional, repeatable.

In Figure 4, we describe how these elements are related to the metadata templates of each data layer.

**Figure 4: DDI Lite v2.1**

In [Appendix 1](#) the complete DDI-Lite template is provided. In [Appendix 2](#) we give a number of examples for the DDI-Lite metadata in XML for data sets in Layer 1 (source corresponding), Layer 2 (release of processed data) and Layer 3 (a data set for analysis). The examples are based on the Historical Sample of the Netherlands (HSN)\(^5\). In this project, a reconstruction of life courses based on population registers and civil records is being created (Appendix 2a). At regular intervals, releases for public use are made, sometimes only consisting of parts of the entire data set (regions or sources). For example, Appendix 2B deals with a release for ‘Persoonskaarten’ records described persons. For particular studies, researchers use a release and modify this, e.g. by adding or modifying variables. Thus, they create a new data set for analysis for which specific documentation is required, to allow for possible verification or replication of the research. An example of metadata accompanying a ‘data set for analysis’ based on an underlying HSN Release is given in Appendix 2C.

---

\(^5\) [www.iisg.nl/hsn/](http://www.iisg.nl/hsn/)
Setting rights on collectively created data sets

In the above, we have assumed that the research team in a collaboratory has pooled the data and has, at some point, transferred the common data set and the various releases to a trusted repository or data archive. The archive provides access to the data sets for reuse by researchers who can enhance and redistribute the data themselves. This assumption may be too idealistic and naive. Several hurdles, associated with data rights, have to be taken.

In principle, everybody who has made a ‘creative’ contribution to a data set has rights on the data he or she contributed. ‘Creative contribution’ is to be seen here as participation in the construction of the data structure, not just data entry. In the case of pooling already existing data from different countries, the (not fully crystallized) international laws on databases have to be taken into account. According to the European Directive regarding the protection of databases (96/9/EC), for data which have been produced or published less than 15 years ago, consent has to be given by the original creators of the data.

To complicate things further, rights on data may belong to the organizations the individual scholars work for. In the Netherlands, however, the individuals who created the data by default own the right to the data. When collaboratories consisting of international teams start pooling data, it is strongly advised that they check and clear the ownership of the rights on the data.

A new situation emerges, when collaboratories jointly create a new data set. This gives more leeway to reach an agreement on how to handle data rights. We advise strongly that – after a period of embargo in which the team can collect and harmonize the data and perform research on them – an open data licence is adopted. In such a licence, rights on the data are waived and data are provided freely to the public, on the condition of proper citation. Although not fully developed yet, the Zero Waiver of Creative Commons (CC0) seems most appropriate to data collected within collaboratories:

"CC0 is a protocol that enables people to WAIVE to the fullest extent possible under applicable copyright law all rights they have and associate with a work so it has no (or minimal) copyright or neighbouring rights restrictions attached to it. To the extent the waiver is not legally effective in any jurisdiction, then the protocol takes the form of a nonexclusive worldwide license to exercise all copyright and neighbouring legal rights in the work".

---

7 http://wiki.creativecommons.org/CCZero
6 Conclusion

In this preliminary effort to draw up practical guidelines for the preservation of research data, we have situated the data collection and archiving processes within the research landscape. We have elaborated a generic layered data model for data collection and archiving based on the experiences of the IISH with data collection collaboratories for Global Labour and Economic History and experiences of the Historical Sample of the Netherlands data center (HSN), which is also a part of the IISH. We have looked into the data curation responsibilities, according to the layered data model and very superficially, explored the data rights issues.

In conclusion, it can be stated that the layered data model is a useful model to distinguish different stages in the data collection process. The dynamics of data collection in online collaboratories illustrate this process best and the analysis has highlighted the iterative approach to data collection, data processing, data analysis and publication of (interim) research results. The research practice drives the requirements for data archiving. The practice also shows that data archiving cannot be conceived as a one-time data transfer transaction or even a linear process. The best guarantee for effective and loss-less preservation of research data can be achieved by interconnecting the data archiving and data collection processes, during the data set release iterations. By depositing data set releases from layer 1 (source corresponding data set) and layer 2 (processed data set) with data archives, we suggest that significant reduction in redundancy can be achieved in the storage and distribution of research data. This can only be achieved by applying rigorous versioning practices to data set releases and by ensuring the referential integrity of data sets, all of which are recognized best practices in the field of data archiving. In order to alleviate the structural costs of data preservation and availability, it would be worthwhile to investigate more thoroughly the effects of these best practices on data redundancy.

The layered data model is defined as a generic model but its usefulness across all scientific disciplines still needs to be demonstrated. For the historical sciences, sharing the source corresponding data layer is crucial to leverage the effort spent by data collection projects. The question now arises how strongly the digitization of cultural heritage and scientific collections and the advance of data mining technologies will accelerate the growth of this layer?

The data layer model also provides room for differentiating between the data rights management and access embargo regimes for the different layers. It could be agreed for example that releases of the source corresponding data layer become open access immediately upon release. This needs further exploration.
Appendix 1. DDI Lite template for research data

The Fields below comprise the DDI Lite ([http://www.ddialliance.org/dtd/](http://www.ddialliance.org/dtd/)), a list of recommended elements of the full DDI, or the fields needed to make the metadata XML usable. Fill out all fields if they are applicable to the research data set being released.

<table>
<thead>
<tr>
<th>Element</th>
<th>Full name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>docDscr</td>
<td>Document Description</td>
<td>The section Document Description provides information on the DDI Document itself. Who created it, version of the marked up document, how to cite it etc.</td>
</tr>
<tr>
<td>Citation</td>
<td>Bibliographic Citation</td>
<td>Bibliographic information includes title information, statement of responsibility, production and distribution information, series and version information, text of a preferred bibliographic citation, and notes (if any).</td>
</tr>
<tr>
<td>titlStmt</td>
<td>Title Statement</td>
<td>Title statement for the marked-up document</td>
</tr>
<tr>
<td>IDno</td>
<td>Identification Number</td>
<td>Unique string or number for the marked up document</td>
</tr>
<tr>
<td>prodStmt</td>
<td>Production statement</td>
<td>Production statement for the marked-up document</td>
</tr>
<tr>
<td>Producer</td>
<td>Producer</td>
<td>The &quot;producer&quot; in the Document Description should be the organization or person that prepared the marked-up document.</td>
</tr>
<tr>
<td>prodDate</td>
<td>Production Date</td>
<td>Date the marked up document was produced</td>
</tr>
<tr>
<td>copyright</td>
<td>Copyright statement</td>
<td>Copyright statement for the marked up document</td>
</tr>
<tr>
<td>software</td>
<td>Software used in Production</td>
<td>Software used to make the marked up document</td>
</tr>
<tr>
<td>verStmt</td>
<td>Version Statement</td>
<td>Version statement for the marked-up document</td>
</tr>
<tr>
<td>Notes</td>
<td>Notes</td>
<td>For clarifying information/annotation on the version</td>
</tr>
<tr>
<td>stdyDscr</td>
<td>Study Description</td>
<td>The Study Description consists of information about the data collection, study, or compilation that the DDI-compliant documentation file describes. This section includes information about how the study should be cited, who collected or compiled the data, who distributes the data, keywords about the content of the data, summary (abstract) of the content of the data, data collection methods and processing, etc.</td>
</tr>
<tr>
<td>citation</td>
<td>Bibliographic Citation</td>
<td>Bibliographic information includes title information, statement of responsibility, production and distribution information, series and version information, text of a preferred bibliographic citation, and notes (if any).</td>
</tr>
<tr>
<td>titlStmt</td>
<td>Title Statement</td>
<td>Title statement for the study</td>
</tr>
<tr>
<td>tti</td>
<td>Title</td>
<td>Full authoritative title of the study</td>
</tr>
<tr>
<td>IDno</td>
<td>Identification Number</td>
<td>Unique string or number for the study</td>
</tr>
<tr>
<td>rspStmt</td>
<td>Responsibility Statement</td>
<td>Responsibility for the creation of the study</td>
</tr>
<tr>
<td>AuthEnty</td>
<td>Authoring Entity/Primary Investigator</td>
<td>The person, corporate body, or agency responsible for the work's substantive and intellectual content. Repeat the element for each author, and use “affiliation” attribute if available. Invert first and last name and use commas.</td>
</tr>
<tr>
<td>othId</td>
<td>Other Identifications /Acknowledgments</td>
<td>Statements of responsibility not recorded in the title and statement of responsibility areas. Indicate here the persons or bodies connected with the work, or significant persons or bodies connected with previous editions and not already named in the description.</td>
</tr>
<tr>
<td>prodStmt</td>
<td>Production Statement</td>
<td>Production statement for the study</td>
</tr>
<tr>
<td>Element</td>
<td>Full name</td>
<td>Description</td>
</tr>
<tr>
<td>----------------------</td>
<td>------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Producer</td>
<td>Producer</td>
<td>The organization or person that prepared the marked-up document.</td>
</tr>
<tr>
<td>prodDate</td>
<td>Production Date</td>
<td>Date the source/data collection/other material(s) were produced (not distributed or archived)</td>
</tr>
<tr>
<td>copyright</td>
<td>Copyright statement</td>
<td>Copyright for data collection</td>
</tr>
<tr>
<td>fundAg</td>
<td>Funding Agency/Sponsor</td>
<td>The source(s) of funds for production of the work</td>
</tr>
<tr>
<td>distStmt</td>
<td>Distributor Statement</td>
<td>Distribution statement for the study</td>
</tr>
<tr>
<td>distrbtr</td>
<td>Distributor</td>
<td>The organization designated by the author or producer to generate copies of the particular work including any necessary editions or revisions.</td>
</tr>
<tr>
<td>distDate</td>
<td>Date of Distribution</td>
<td>Date that the work was made available for distribution/presentation</td>
</tr>
<tr>
<td>serStmt</td>
<td>Series Statement</td>
<td>Series statement for the study</td>
</tr>
<tr>
<td>serName</td>
<td>Series Name</td>
<td>The name of the series to which the work belongs.</td>
</tr>
<tr>
<td>verStmt</td>
<td>Version Statement</td>
<td>Version statement for the study</td>
</tr>
<tr>
<td>version</td>
<td>Version</td>
<td>Also known as release or edition</td>
</tr>
<tr>
<td>notes</td>
<td>Notes</td>
<td>For clarifying information/annotation regarding the version</td>
</tr>
<tr>
<td>biblCit</td>
<td>Bibliographic Citation</td>
<td>Complete bibliographic reference containing all of the standard elements of a citation that can be used to cite the study.</td>
</tr>
<tr>
<td>stdyInfo</td>
<td>Study Scope</td>
<td>This section contains information about the data collection's scope across several dimensions, including substantive content, geography, and time.</td>
</tr>
<tr>
<td>subject</td>
<td>Subject</td>
<td>Subject information describing the data collection's intellectual content</td>
</tr>
<tr>
<td>keyword</td>
<td>Keyword</td>
<td>Words or phrases that describe salient aspects of a data collection's content</td>
</tr>
<tr>
<td>abstract</td>
<td>Abstract</td>
<td>An unformatted summary describing the purpose, nature, and scope of the data collection, special characteristics of its contents, major subject areas covered, and what questions the PIs attempted to answer when they conducted the study. A listing of major variables in the study is important here.</td>
</tr>
<tr>
<td>sumDscr</td>
<td>Summary Data Description</td>
<td>Information about a study's chronological and geographic coverage and unit of analysis</td>
</tr>
<tr>
<td>timePrd</td>
<td>Time Period Covered</td>
<td>The time period to which the data refer.</td>
</tr>
<tr>
<td>collDate</td>
<td>Date of Collection</td>
<td>Contains the date(s) when the data were collected</td>
</tr>
<tr>
<td>nation</td>
<td>Country</td>
<td>Indicates the country or countries covered in the file</td>
</tr>
<tr>
<td>geogCover</td>
<td>Geographic Coverage</td>
<td>Information on the geographic coverage of the data</td>
</tr>
<tr>
<td>anlyUnit</td>
<td>Unit of Analysis</td>
<td>Basic unit of analysis or observation that the file describes: individuals, families/households, groups, institutions/organizations, administrative units</td>
</tr>
<tr>
<td>universe</td>
<td>Universe</td>
<td>Group of persons or other elements that are the object of research and to which any analytic results refer.</td>
</tr>
<tr>
<td>dataKind</td>
<td>Kind of Data</td>
<td>The type of data included in the file: survey data, census/enumeration data, aggregate data, clinical data, event/transaction data, program source code, machine-readable text, administrative records data,</td>
</tr>
<tr>
<td>notes</td>
<td>Notes</td>
<td>For clarifying information/annotation regarding the data collection</td>
</tr>
<tr>
<td>method</td>
<td>Methodology and Processing</td>
<td>This section describes the methodology and processing involved in a data collection</td>
</tr>
<tr>
<td>dataColl</td>
<td>Data Collection Methodology</td>
<td>Information about the methodology employed in a data collection</td>
</tr>
<tr>
<td>timeMeth</td>
<td>Time Method</td>
<td>The time method or time dimension of the data collection</td>
</tr>
<tr>
<td>dataCollector</td>
<td>Data Collector</td>
<td>This refers to the entity collecting the data</td>
</tr>
</tbody>
</table>

30
<table>
<thead>
<tr>
<th>Element</th>
<th>Full name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>sampProc</td>
<td>Sampling Procedure</td>
<td>The type of sample and sample design used to select the survey respondents to represent the population.</td>
</tr>
<tr>
<td>collMode</td>
<td>Mode of Data Collection</td>
<td>The method used to collect the data; instrumentation characteristics.</td>
</tr>
<tr>
<td>sources</td>
<td>Sources Statement</td>
<td>Description of sources used for the data collection.</td>
</tr>
<tr>
<td>weight</td>
<td>Weighting</td>
<td>The use of sampling procedures may make it necessary to apply weights to produce accurate statistical results. Describe here the criteria for using weights in analysis of a collection. If a weighting formula or coefficient was developed, provide this formula, define its elements, and indicate how the formula is applied to data.</td>
</tr>
<tr>
<td>cleanOps</td>
<td>Cleaning Operations</td>
<td>Methods used to &quot;clean&quot; the data collection, e.g., consistency checking, wildcode checking, etc.</td>
</tr>
<tr>
<td>notes</td>
<td>Notes</td>
<td>For clarifying information/annotation regarding the method</td>
</tr>
<tr>
<td>dataAccs</td>
<td>Data Access</td>
<td>This section describes access conditions and terms of use for the data collection.</td>
</tr>
<tr>
<td>setAvail</td>
<td>Data Set Availability</td>
<td>Information on availability and storage of the collection.</td>
</tr>
<tr>
<td>coilSize</td>
<td>Extent of Collection</td>
<td>Summarizes the number of physical files that exist in a collection, recording the number of files that contain data and noting whether the collection contains machine-readable documentation and/or other supplementary files and information such as data dictionaries, data definition statements, or data collection instruments.</td>
</tr>
<tr>
<td>fileQnty</td>
<td>Number of Files</td>
<td>Total number of physical files associated with a collection.</td>
</tr>
<tr>
<td>useStmt</td>
<td>Use Statement</td>
<td>Information on terms of use for the data collection.</td>
</tr>
<tr>
<td>restrctn</td>
<td>Restrictions</td>
<td>Any restrictions on access to or use of the collection such as privacy certification or distribution restrictions should be indicated here.</td>
</tr>
<tr>
<td>notes</td>
<td>Notes</td>
<td>For clarifying information/annotation regarding the DataAccess</td>
</tr>
<tr>
<td>fileDscr</td>
<td>Data Files Description</td>
<td>Information about the data file(s) that comprises a collection. This section can be repeated for collections with multiple files.</td>
</tr>
<tr>
<td>fileTxt</td>
<td>File-by-File Description</td>
<td>Information about the data file.</td>
</tr>
<tr>
<td>fileName</td>
<td>File Name</td>
<td>Contains a short title that will be used to distinguish a particular file/part from other files/parts in the data collection.</td>
</tr>
<tr>
<td>dimensns</td>
<td>File Dimensions</td>
<td>Dimensions of the overall file.</td>
</tr>
<tr>
<td>caseQnty</td>
<td>Number of cases / Record Quantity</td>
<td>Number of cases or observations.</td>
</tr>
<tr>
<td>varQnty</td>
<td>Number of variables per record</td>
<td>Number of variables.</td>
</tr>
<tr>
<td>logRecL</td>
<td>Record Length / Logical Record Length</td>
<td>Logical record length, i.e., number of characters of data in the record.</td>
</tr>
<tr>
<td>recPrCas</td>
<td>Records per Case</td>
<td>Records per case in the file</td>
</tr>
<tr>
<td>fileType</td>
<td>Type of File</td>
<td>Types of data files</td>
</tr>
<tr>
<td>notes</td>
<td>Notes</td>
<td>For clarifying information/annotation regarding the fileDscr</td>
</tr>
<tr>
<td>dataDscr</td>
<td>Variable Description</td>
<td>Description of variables</td>
</tr>
<tr>
<td>varGrp</td>
<td>Variable Group</td>
<td>A group of variables that may share a common subject, arise from the interpretation of a single question, or are linked by some other factor.</td>
</tr>
<tr>
<td>labl</td>
<td>Label</td>
<td>Description of the variable</td>
</tr>
<tr>
<td>Var</td>
<td>Variable</td>
<td>This element describes all of the features of a single variable in a social science data file.</td>
</tr>
<tr>
<td>location</td>
<td>Location</td>
<td>StartPos, EndPos, width, RecSegNo, fileid, locMap</td>
</tr>
<tr>
<td>Qstn</td>
<td>Question</td>
<td>Research Question</td>
</tr>
<tr>
<td>qstnLit</td>
<td>Literal Question</td>
<td>Text of the actual, literal question asked</td>
</tr>
<tr>
<td>Element</td>
<td>Full name</td>
<td>Description</td>
</tr>
<tr>
<td>---------------</td>
<td>----------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Invalrng</td>
<td>Range of Invalid Data Values</td>
<td>Values for a particular variable that represent missing data</td>
</tr>
<tr>
<td>Notes</td>
<td>Range Key</td>
<td>This element permits a listing of the category values and labels.</td>
</tr>
<tr>
<td>universe</td>
<td>Universe</td>
<td>The group of persons or other elements that are the object of research and to which any analytic results refer.</td>
</tr>
<tr>
<td>sumStat</td>
<td>Summary Statistics</td>
<td>One or more statistical measures that describe the responses to a particular variable and may include one or more standard summaries, e.g., minimum and maximum values, median, mode, etc</td>
</tr>
<tr>
<td>Txt</td>
<td>Descriptive Text</td>
<td>Lengthier description of variables</td>
</tr>
<tr>
<td>catgryGrp</td>
<td>Category Group</td>
<td>A description of response categories that might be grouped together.</td>
</tr>
<tr>
<td>Labl</td>
<td>Label</td>
<td>Description of the category grp</td>
</tr>
<tr>
<td>catStat</td>
<td>Category Level Statistic</td>
<td>May include frequencies, percentages, or crosstabulation results.</td>
</tr>
<tr>
<td>Concept</td>
<td>Concept</td>
<td>A description of a particular response.</td>
</tr>
<tr>
<td>Derivation</td>
<td>Derivation</td>
<td>Used only in the case of a derived variable, this element provides both a description of how the derivation was performed and the command used to generate the derived variable, as well as a specification of the other variables in the study used to generate the derivation</td>
</tr>
<tr>
<td>Drvdesc</td>
<td>Derivation Description</td>
<td>A textual description of the way in which this variable was derived.</td>
</tr>
<tr>
<td>varFormat</td>
<td>Variable Format</td>
<td>The technical format of the variable in question</td>
</tr>
<tr>
<td>notes</td>
<td>Notes</td>
<td>For clarifying information/annotation regarding the dataDscr</td>
</tr>
<tr>
<td>otherMat</td>
<td>Other Study-Related Materials</td>
<td>This section allows for the inclusion of other materials that are related to the study as identified and labeled by the DTD/Schema users (encoders). The materials may be entered as PCDATA (ASCII text) directly into the document (through use of the &quot;txt&quot; element). This section may also serve as a &quot;container&quot; for other electronic materials such as setup files by providing a brief description of the study-related materials accompanied by the attributes &quot;type&quot; and &quot;level&quot; defining the material further. The &quot;URI&quot; attribute may be used to indicate the location of the other study-related materials.</td>
</tr>
<tr>
<td>Notes</td>
<td>Notes</td>
<td>For clarifying information/annotation regarding the</td>
</tr>
<tr>
<td>Table</td>
<td>Table</td>
<td>Table</td>
</tr>
<tr>
<td>citation</td>
<td>Bibliographic Citation</td>
<td>Bibliographic information includes title information, statement of responsibility, production and distribution information, series and version information, text of a preferred bibliographic citation, and notes (if any).</td>
</tr>
<tr>
<td>titlStmt</td>
<td>Title Statement</td>
<td>Title statement for the other material</td>
</tr>
<tr>
<td>titl</td>
<td>Title</td>
<td>Full authoritative title of the other material</td>
</tr>
</tbody>
</table>
Appendix 2a. Example of XML metadata for a source-corresponding data set

Example of XML metadata for a source-corresponding dataset (page 1)

```xml
<xml version="1.0" encoding="utf-8">
<codeBook xmlns="http://www.icpsr.umich.edu/DDI">
<!-- Example DDI for processed Dataset - Layer 1-->
<!-- Document Description -->
<docDscr>
  <citation>
    <titlStmt>
      <titl>Historical Sample of the Netherlands, 1817 - 1922: processed Dataset 1.0</titl>
      <IDNo agency="HSN">1234</IDNo>
    </titlStmt>
    <prodStmt>
      <producer abbr="HSN">Historical Sample of the Netherlands</producer>
      <prodDate date="2007-10-01">October 10, 2007</prodDate>
      <copyright>copyright: HSN</copyright>
      <software version="5.0">XMetal Author</software>
    </prodStmt>
    <verStmt>
      <version type="Dataset" date="2007-10-01">First version</version>
      <notes>This is the metadata document for the HSN processed Dataset 1.0.</notes>
    </verStmt>
  </citation>
  <notes>The source codebook was produced from original hardcopy material</notes>
</docDscr>

<!-- Study Description -->
<stdyDscr>
  <citation>
    <titlStmt>
      <titl>Historical Sample of the Netherlands, 1817 - 1922: processed Dataset 1.0</titl>
      <IDNo agency="HSN">1234</IDNo>
    </titlStmt>
    <rspStmt>
      <AuthEnty>Mandemakers, Kees</AuthEnty>
      <othId role="contributor" affiliation="HSN">Hendrien Emmers</othId>
      <othId role="contributor" affiliation="HSN">Hendrien Emmers</othId>
      <othId role="contributor" affiliation="HSN">Henk van der Gaauw</othId>
      <othId role="contributor" affiliation="HSN">Kees Mandemakers</othId>
    </rspStmt>
    <prodStmt>
      <producer affiliation="HSN">Kees Mandemakers</producer>
      <prodDate date="2007-10-01">October 10, 2007</prodDate>
      <copyright>HSN</copyright>
      <fundAg>Dutch foundation of scientific research (NWO)</fundAg>
    </prodStmt>
    <distStmt>
      <distrbtr abbr="HSN">Historical Sample of the Netherlands</distrbtr>
      <distDate date="2007-10-01">October 10, 2007</distDate>
    </distStmt>
    <verStmt>
      <version type="Dataset" date="2007-10-01">First version</version>
      <notes>This is the first version of the processed dataset.</notes>
    </verStmt>
    <biblCit>Mandemakers, Kees., Historical Sample of the Netherlands, 1817 - 1922: processed Dataset 1.0 [codebook] Conducted by HSN Organisation. Amsterdam, 2007</biblCit>
  </citation>
  <stdyInfo>
    <subject>Dutch population registers and censuses</subject>
    <keywords>census</keywords>
    <fileTxt>The data base consists of seven files, generated from the population registers and the birth certificates, all related to each other. </fileTxt>
    <fileDscr ID="HNS" URI="www.iisg.nl/repository/oreresearchdata/files/1234/>
    <fileTxt>
      <fileId>1234</fileId>HSN processed</fileId>
      <fileCont>The core of the HSN data set is the research person (RP). This has implications for every HSN program used for entering or processing data. In principle, all other individuals and corresponding attributes are defined in relation to the research person. Unlike the population register, that is defining every person in relation to the head of the household. An RP can be a head of the household, or just a member of the household. During the course of its life, an RP can be member of different households, with different heads.</fileTxt>
  </stdyInfo>
</stdyDscr>
</codeBook>
</xml>
```
Out of the information from the population registers, the output program creates a data base, being made up of five tables. Technically, the research unit of these tables is the household. However, if there’s not a research person, no household is being processed. During the course of processing data, the birth certificate of the RP is also included.

In addition to the above mentioned five tables, one table with ‘meta information’ is created: ‘BEVOP’, summarizing information about the RP. In this abridged version of the manual, a seventh output table has been skipped. This table is only relevant when research persons are related to each other, e.g. father and son (1st and 2nd generation).

HSN distinguishes households, individuals and household blocks. A household block is a collection of individuals which, according to source data, is to be considered a unit. The minimum number of individuals in a household block is one. Usually every single household block starts with the head (of the household) at the first line. A household block may consist of more than one household, for example when the head of the household died. In case of registers of single persons, the application considers the single person to be the head of the household.

The HSN application combines different household blocks to households. All redundant information is being removed during this process. Every household has its own head. In the event, a head of a household died, and the population register did not define a new head, the HSN application itself appoints a new one. In case new households are being formed, as a result of death or departure of the original head, these new households are nothing but artificial constructions. However, such constructions are necessary, because the relations in the population register are defined or have to be defined to the new head of the household, especially after a change from one register to the other one. Households being constructed this way, being part of the same household blocks, are identified in the HSN application by a special variable. This makes it possible to reconstruct the original situation, or to combine different related households, e.g. the household of an original head with that of his widow.

As a matter of principle, members of a household are only those observed for the time being the RP is part of the household. Newly created households after a definite departure of the RP, are not part of the output.

Sometimes, the RP never was found in a population register, for example in case of an early death. However, not being able to find an RP, does not always mean the RP was not registered. For example, the RP could have been registered at other places, with the parents of the mother or with other family members, without any clue from birth certificates, death certificates, address lists or migration data. In all these cases in the household where the mother was registered at the time of birth, a ‘fake RP’ was constructed. In case a household moved to another community, and moved back later on, ‘gaps’ are possible, if the data of that other community were not entered. These gaps can be deduced from individual departure and return dates.

In principle, all different population registers are regarded as one source. So, the application links all registrations from the same head, independent from actual residential areas in The Netherlands. In this way, the observation period of the household becomes equal to the observation period of the head of the household.

For the period after 1900, one should realise that ‘households’ are usually no more than nuclear families (father, mother and children). All in all, in literature names like households, household blocks and families are not very clear defined. For HSN definitions, see further appendix A.

In the output, all dates are given three variables: day, month and year. However, in population registers dates of registration, departure and arrival are often lacking. These dates are automatically estimated. Estimated dates are identified by day numbers higher than 31. Dates entered as day number, month number and year number, are also converted into date fields, structured like ‘mm/dd/yy’. In cases of day numbers higher than 31, during the conversion a choice was made for the first, the middle or the end of the month, depending on the situation. An exception is code ‘40’, which stands for a ‘declared date’ which is a date for a specific situation at a certain time (for example ‘being married’ at the start of a registration). These types of ‘dates’ are not converted into date fields.

Often programs used for analyzing data, like SPSS, are only able to handle limited lengths of variable names. Therefore, variable names used by HSN applications, are not exceeding eight characters.
Appendix 2b. Example of XML metadata for a release of a processed data set

```xml
<xml version="1.0" encoding="utf-8">
<codebook xmlns="http://www.icpsr.umich.edu/DDI">
<!-- Example DDI for Release - Layer 2-->
<!-- Document Description -->
<docDscr>
  <citation>
    <IDNo agency="HSN">1234</IDNo>
    <prodStmt>
      <producer abbr="HSN">Historical Sample of the Netherlands</producer>
      <prodDate date="2008-10-01">October 10, 2008</prodDate>
      <copyright>copyright: HSN</copyright>
      <software version="5.0">XMetal Author</software>
    </prodStmt>
    <version type="release" date="2008-10-01">First version</version>
    <notes>This is the metadata document for the HSN Release: Persoonskaarten Release 2008.1</notes>
  </citation>
  <notes>The source codebook was produced from original hardcopy material</notes>
</docDscr>

<!-- Study Description -->
<stdyDscr>
  <citation>
    <IDNo agency="HSN">1234</IDNo>
    <rspStmt>
      <AuthEnty>Mandemakers, Kees</AuthEnty>
      <othId role="contributor" affiliation="HSN">Hendrie Emmers</othId>
      <othId role="contributor" affiliation="HSN">Behice Gül</othId>
      <othId role="contributor" affiliation="HSN">Henk van der Gaauw</othId>
      <othId role="contributor" affiliation="HSN">Kees Mandemakers</othId>
    </rspStmt>
    <prodStmt>
      <producer affiliation="HSN">Kees Mandemakers</producer>
      <prodDate date="2008-10-01">October 10, 2008</prodDate>
      <copyright>HSN</copyright>
      <fundAg>Dutch foundation of scientific research (NWO)</fundAg>
    </prodStmt>
    <distStmt>
      <distrbtr abbr="HSN">Historical Sample of the Netherlands</distrbtr>
      <distDate date="2008-10-01">October 10, 2008</distDate>
    </distStmt>
    <verStmt>
      <version type="release" date="2007-10-01">First version</version>
      <notes>This is the first version of the Release: Persoonskaarten Release 2008.1</notes>
    </verStmt>
  </citation>
</stdyDscr>
</codebook>
</xml>`
## Appendix 2c. Example of XML metadata for a Dataset for analysis

### Example of XML metadata for a Dataset for analysis (page 1)

```xml
<xml version="1.0" encoding="utf-8">
<codebook xmlns="http://www.icpsr.umich.edu/DDI">
  <!-- Example DDI for Dataset for Analysis Release - Layer 3-->
  <docDscr>
    <citation>
      <titl>Historical Sample of the Netherlands, 1817 - 1922: Continuity and Change: Dataset for Analysis 2009.1</titl>
      <IDM agency="HSN">1234</IDM>
    </citation>
    <prodStmt>
      <producer abbr="HSN">Historical Sample of the Netherlands</producer>
      <prodDate date="2009-02-01">February 1, 2009</prodDate>
      <copyright>Historical Sample of the Netherlands</copyright>
      <software version="5.0">XMetal Author</software>
    </prodStmt>
    <verStmt>
      <version type="release" date="2009-02-01">First version</version>
      <notes>This is the metadata document for the HSN Continuity and Change: Dataset for Analysis 2009.1</notes>
    </verStmt>
  </docDscr>
  <stdyDscr>
    <citation>
      <titl>Historical Sample of the Netherlands, 1817 - 1922: Continuity and Change: Dataset for Analysis 2009.1</titl>
      <IDM agency="HSN">1234</IDM>
      <respStmt>
        <AuthEnty>Kok, Jan</AuthEnty>
        <othId role="contributor" affiliation="HSN">Mandemakers, Kees</othId>
      </respStmt>
      <prodStmt>
        <producer affiliation="HSN">Jan Kok</producer>
        <prodDate date="2009-02-01">February 1, 2009</prodDate>
        <copyright>Historical Sample of the Netherlands</copyright>
        <fundAg>Dutch foundation of scientific research (NWO)</fundAg>
      </prodStmt>
      <distStmt>
        <distrbtr abbr="HSN">Historical Sample of the Netherlands</distrbtr>
        <distDate date="2009-02-01">February 1, 2009</distDate>
      </distStmt>
      <verStmt>
        <version type="release" date="2009-02-01">First version</version>
        <notes>This is the first version of the Continuity and Change: Dataset for Analysis 2009.1</notes>
      </verStmt>
      <biblCit>Kok, Jan., Historical Sample of the Netherlands, 1817 - 1922: Continuity and Change: Dataset for Analysis 2009.1 [codebook] Conducted by HSN Organisation. Amsterdam, 2008</biblCit>
    </citation>
  </stdyDscr>
  <dataDscr>
    <varGrp>
      <labl>Idnr</labl>
      <txt>Identificatienummer, sleutel naar BEV0P, HSN identification number of the research person (RP)</txt>
    </varGrp>
    <varGrp>
      <labl>Huishnr</labl>
      <txt>Huishoudnummer, sleutel naar BEVHUISH</txt>
    </varGrp>
    <varGrp>
      <labl>Idnrhsp</labl>
      <txt>Uniek nummer verwijzend naar de oorspronkelijke records in de bestanden Bevdynap en bevesttstt van onderhavige records (nummer van het hoofd van het huishouden)</txt>
    </varGrp>
    <varGrp>
      <labl>Idnrhs_nw</labl>
      <txt>Nieuw huishoudnummer (vergelijken met huisnr extra digits vanwege splitsing huishoudens na verandering burgerlijke staat van het hoofd en na splitsing van nucleaire huishoudens in met en zonder kinderen)</txt>
    </varGrp>
    <varGrp>
      <labl>Pers0P_v</labl>
      <txt>Persoonsnummer van het hoofd van het huishouden</txt>
    </varGrp>
  </dataDscr>
</codebook>
</xml>
```
Example of XML metadata for a Dataset for analysis (page 2)

- A: OP is alleenwonend
- B: OP is inwonend als niet-verwant
- C: OP is inwonend als niet-verwant (woont OP ongehuwd samen -> wordt inwonend als niet-verwant)
- D: OP is zelf hoofd
- E: OP is gehuwd met het hoofd
- F: OP is kind van het hoofd (inclusief stiefkinderen)
- G: OP is ouder van het hoofd (inclusief stiefouders)
- H: OP is kleinkind van het hoofd