Arhiv družboslovnih podatkov

# PRIPRAVA PODATKOV IN DOKUMENTACIJE

STRATEGIJA PRIPRAVE PODATKOVNIH DATOTEK
in DOKUMENTACIJE

Irena **Vipavc Brvar**

Oktober 2013

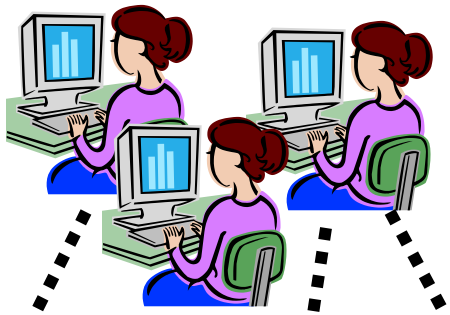
# Dokumentacija

Priprava kakovostne tehnične dokumentacije ali tako imenovane kodirne knjige je lahko časovno izredno zahtevna. Tukaj govorimo tako o ustrezni pripravi podatkovne datoteke (mikro podatkov) kot spremljajočega gradiva (metapodatkov)

Podrobni vodiči so dostopni na spletnih straneh angleškega (1, 2) in ameriškega arhiva. V nadaljevanju pa sledi nekaj krajših napotkov.

Primarni raziskovalec

Sodelujoči



Raziskovalno osebje

<DDI 3.0>
Purpose
Concepts
Universe
Geography
People/Orgs

+

<DDI 3.0>
Funding
Revisions

+

<DDI 3.0>
Questions
Instrument

+

<DDI 3.0>
Data Collection
Data Processing

+

<DDI 3.0>
Variables
Physical Stores

Data

$

€ £

Podan predlog projekta

Publikacije

Arhiv/ Repozitorij

# STRATEGIJA

## KAJ HRANITI ?
## KAKO HRANITI ?

Pomembno:

- Pripravi varnostne kopije pomembnih datotek
- Nudi ustrezno pretvorbo formatov
- Zagotovi avtentičnost in kontroliran dostop do hranjenih in delovnih datotek
- Zagotovi kontrolo verzij
- Nudi primerni pomnilnik
- Zagotovi varnost

# PRIPRAVA PODATKOVNE DATOTEKE 1

-Tvorimo vnašalnik, ki bo omogočal preverjanje vrednosti vnosov.

-Programsko izvozimo celotne opise spremenljivk iz vprašalnika v podatkovno datoteko (se izognemo številnim napakam, ki bi nastala pri ročnem vnosu)

-Preverimo kvaliteto vnosa (naključnih 10%, prvih in zadnjih 5%)

-Ločimo osebe, ki kodirajo z osebami, ki vnašajo podatke. Zahtevna kodiranja (poklicne kode) opravijo izkušeni posamezniki. Kar lahko naredi računalnik ne dejmo sami.

# PRIPRAVA PODATKOVNE DATOTEKE 2

-Imena spremenljivk

-Opisi spremenljivk

-Združevanje skupin spremenljivk

-Kode in kodiranje

-Manjkajoče vrednost

# MANJKAJOČE VREDNOSTI

Če je le mogoče **ne uporabljamo praznih polj** ampak definiramo vrednosti. Le v primeru če manjka večina odgovorov pri enoti/ ali odgovor na spremenljivko v ponovljivi raziskavi.

Pomembno je razlikovati med **različnimi oblikami** manjkajočih vrednosti:

- Zavnitev /neodgovor
- Ne vem
- Napaka v postopku
- Neustrezen

**Imputacija** manjkajočih vrednosti! (original + izpeljana)

**Koda** za manjkajoče vrednosti (97,98,99 / -1, -9) – problem s št. mest

# ČIŠČENJE PODAKTOVNE DATOTEKE

**Uporaba sintakse! – sledljivost sprememb**

```
GET FILE='pb1101.sav_vaja'
            /drop= [seznam spremenljivk, ki jih takoj na začetku izločim iz datoteke].

MISSING VALUES   q1 q2 q481b w q695b q695a q702 q696 q704 (3).

VARIABLE LABELS omr 'Omrezna skupina'.

VALUE LABELS omr
   1 '01 - Ljubljana'
   2 '02 - Maribor, Murska Sobota, Ravne na Koroskem'
   3 '03 - Celje, Trbovlje'
   4 '04 - Kranj'
   5 '05 - Koper, Nova Gorica, Postojna'
   7 '07 - Novo Mesto, Krsko'.

EXECUTE .

format wskup (F1.0).

COMPUTE anketa = 1 .
EXECUTE .
COMPUTE anketa = DATE.DMY(15,1,2004) .
EXECUTE .
formats anketa(moyr8).

SAVE  OUTFILE='\pb0501.sav'
            keep= [seznam spremenljivk, ki jih ohranim; v vrstenm redu kot ga želim v datoteki] .
```

# VERZIRANJE IN AVTENTIFIKACIJA

-Odločimo se koliko in katere verzije bomo hranili, kako dolgo in kako naj organiziramo verziranje.

-Enolično poimenovanje z uporabo sistematičnega principa.

-Hranimo verzijo in status datoteke, npr osnutek, vmesni dokument, končna verzija, interni dokument.

-Dokumentiramo spremembe med verzijami datotek.

-Dokumentiramo povezavo med členi ali datotekami, kadar je to potrebno.

-Dokument o lokacijah hrambe datotek, kadar so te na različnih mestih.

-Redno sinhroniziramo datoteke na različnih lokacijah.

-Ohranimo eno glavno datoteko v ustreznem formatu, da se izognemo problemom paralelnega dela na datoteki.

-Indetificiramo eno lokacijo za hrambo glavnih in mejnih verzij.

## Koristi doslednega označevanje datotek, so:

-Podatkovne datoteke so razlikujejo med seboj.
-Ustrezno poimenovanje datotek preprečuje zmedo, ko več ljudi dela na skupnih datotekah.
-Podatkovne datoteke je lažje iskati.
-Podatkovne datoteke lahko prikliče ne le ustvarjalec, ampak tudi drugi uporabniki.
-Podatkovne datoteke lahko razporedimo v logično zaporedje.
-Podatkovnih datotek ni mogoče pomotoma prepisati ali izbrisati.
-Mogoče je prepoznati različne različice datotek.
-Če se podatkovne datoteke preseli v drugo platformo za shranjevanje bodo njihova imena ohranila koristne informacije.

## Koristi doslednega označevanje datotek, so:

Pri imenovanju in označevanju datotek raziskav je potrebno misliti predvsem na naslednje tri kriterije:

- Organizacija - pomembno za prihodnji dostop do gradiv in nj. priklic

- Okvir - to lahko vključuje vsebinsko specifične ali opisne informacije, neodvisno od kje so podatki shranjeni

- Doslednost - izberite poimenovanje in zagotovite, da se pravila sledijo in da se sistematično vključujejo iste informacije (kot so datum in čas), v istem vrstnem redu (npr. DDMMLLLL).

# DOKUMENTIRAJ PODATKE Z UPORABO STRUKTIRANIH OBLIK

Večina CESSDA arhivov danes spodbuja pripravo strukturirane dokumentacije v standardu DDI.

Tako zapisane informacije omogočajo **enostaven izpis v več oblikah** in prenos med različnimi institucijami, ki dokumentacijo uporabljajo v nadaljnjem procesu.

Dokumentacijo poimenujmo **Metapodatki**.

Metapodatke lahko definiramo kot vse informacije potrebne za obveščanje in procesiranje statističnih struktur (Grossmann v Vipavc in Klep, 2003).

# STANDARDI

-**DDI (Data Documentation Initiative )**

-MARC, Dublin Core (bibliographic standards)

SDMX (Statistical Data and Metadata Exchange)

-ISO 11179 (Metadata Registries)

-FGDC (Digital Geospatial Metadata)

-ISO 19115 (Geographic Information Metadata)

- PREMIS (Preservation Metadata), METS (Metadata Encoding and Transmission)

# Enostaven Dublin Core (15 elementov)

Title
Creator
Subject
Description
Publisher
Contributor
Date
Type

Format
Identifier
Source
Language
Relation
Coverage
Rights

# OSNOVNA RAZDELITEV DDI ELEMENTOV

DDI vključuje sledeče pomembne elemente

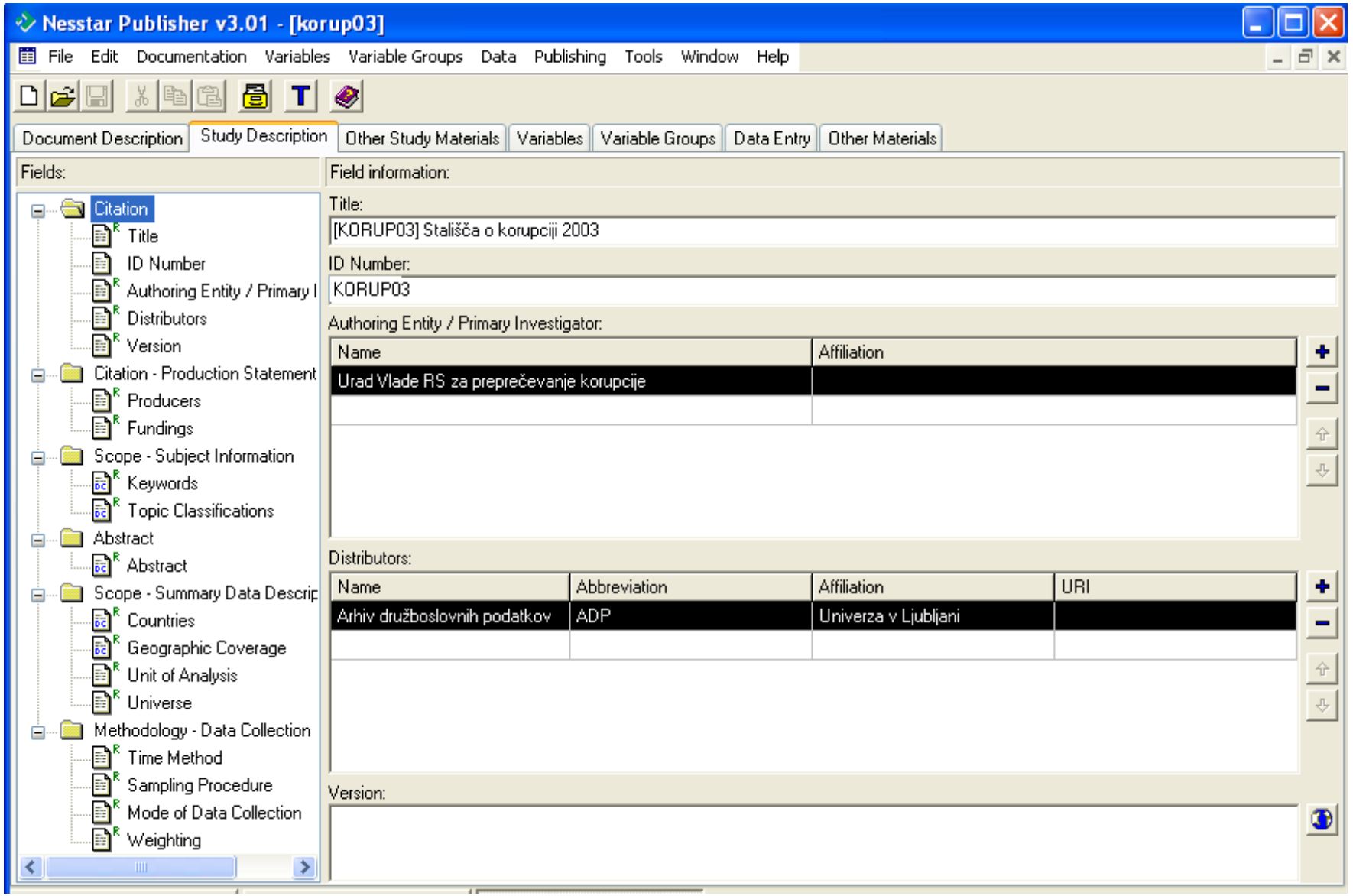
- Opis dokumenta
- Opis raziskave
- Opis podatkovne datoteke
- Opis spremenljivke
- Opis dodatnih in zunanjih, povezanih gradiv

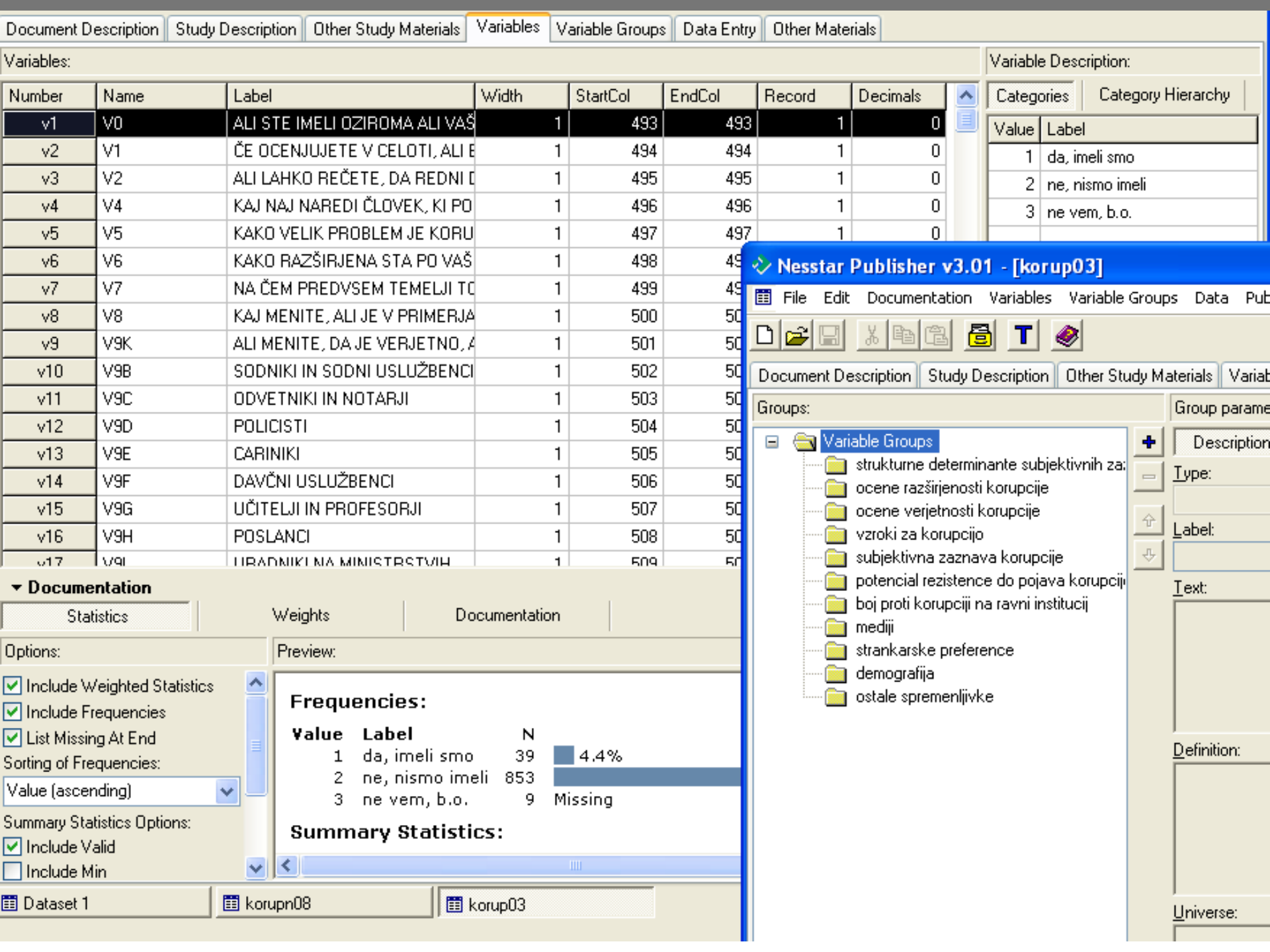
# DDI

| | | |
|---|---|---|
| titl | Title | Naslov |
| IDNo* | Identification Number | Identifikacijska številka |
| prodDate* | Date of Production | Datum izdelave |
| titl | Title | Naslov |
| parTitl* | Paralel title | Vzporedni naslov |
| IDNo* | Identification Number | Identifikacijska številka |
| AuthEnty* | Authoring Entity | Nosilec(ka) raziskave |
| distrbtr* | Distributor | Distribucija podatkov |
| serName* | Series Name | Ime serije |
| version? | Version | Verzija |
| verResp? | Version Responsibility | Odgovornost za verzijo |
| notes* | Notes | Opombe |

# DDI

| | | |
|---|---|---|
| abstract* | Abstract | Povzetek |
| timePrd* | Time Period | Časovno pokritje |
| collDate* | Date of Collection | Datum zbiranja podatkov |
| nation* | Country | Država |
| weight* | Weighting | Uteževanje |
| fileName? | File | Ime podatkovne datoteke |
| var* | Variables | Spremenljivke |
| labl* | Labeles | Labele |


Nesstar Publisher v3.01 - [korup03]
File Edit Documentation Variables Variable Groups Data Publishing Tools Window Help
Document Description
Study Description
Other Study Materials
Variables
Variable Groups
Data Entry
Other Materials
Fields:
Citation
Title
ID Number
Authoring Entity / Primary I
Distributors
Version
Citation - Production Statement
Producers
Fundings
Scope - Subject Information
Keywords
Topic Classifications
Abstract
Abstract
Scope - Summary Data Descrip
Countries
Geographic Coverage
Unit of Analysis
Universe
Methodology - Data Collection
Time Method
Sampling Procedure
Mode of Data Collection
Weighting
Field information:
Title:
[KORUP03] Stališča o korupciji 2003
ID Number:
KORUP03
Authoring Entity / Primary Investigator:
Name
Affiliation
Urad Vlade RS za preprečevanje korupcije
Distributors:
Name
Abbreviation
Affiliation
URI
Arhiv družboslovnih podatkov
ADP
Univerza v Ljubljani
Version:


Document Description
Study Description
Other Study Materials
Variables
Variable Groups
Data Entry
Other Materials
Variables:
Variable Description:
Number
Name
Label
Width
StartCol
EndCol
Record
Decimals
v1 V0 ALI STE IMELI OZIROMA ALI VAŠ 1 493 493 1 0
v2 V1 ČE OCENJUJETE V CELOTI, ALI E 1 494 494 1 0
v3 V2 ALI LAHKO REČETE, DA REDNI D 1 495 495 1 0
v4 V4 KAJ NAJ NAREDI ČLOVEK, KI PO 1 496 496 1 0
v5 V5 KAKO VELIK PROBLEM JE KORU 1 497 497 1 0
v6 V6 KAKO RAZŠIRJENA STA PO VAŠ 1 498
v7 V7 NA ČEM PREDVSEM TEMELJI TO 1 499
v8 V8 KAJ MENITE, ALI JE V PRIMERJA 1 500
v9 V9K ALI MENITE, DA JE VERJETNO, A 1 501
v10 V9B SODNIKI IN SODNI USLUŽBENCI 1 502
v11 V9C ODVETNIKI IN NOTARJI 1 503
v12 V9D POLICISTI 1 504
v13 V9E CARINIKI 1 505
v14 V9F DAVČNI USLUŽBENCI 1 506
v15 V9G UČITELJI IN PROFESORJI 1 507
v16 V9H POSLANCI 1 508
Categories
Category Hierarchy
Value
Label
1 da, imeli smo
2 ne, nismo imeli
3 ne vem, b.o.
Nesstar Publisher v3.01 - [korup03]
File
Edit
Documentation
Variables
Variable Groups
Data
Pub
Document Description
Study Description
Other Study Materials
Variab
Groups:
Group parame
Variable Groups
strukturne determinante subjektivnih za:
ocene razširjenosti korupcije
ocene verjetnosti korupcije
vzroki za korupcijo
subjektivna zaznava korupcije
potencial rezistence do pojava korupcij
boj proti korupciji na ravni institucij
mediji
strankarske preference
demografija
ostale spremenljivke
Description
Type:
Label:
Text:
Definition:
Universe:
Documentation
Statistics
Weights
Documentation
Options:
Preview:
Include Weighted Statistics
Include Frequencies
List Missing At End
Sorting of Frequencies:
Value (ascending)
Summary Statistics Options:
Include Valid
Include Min
Frequencies:
Value Label N
1 da, imeli smo 39 4.4%
2 ne, nismo imeli 853
3 ne vem, b.o. 9 Missing
Summary Statistics:
Dataset 1
korupn08
korup03

# DDI 3.0 Life Cycle Orientation

DDI 3.0 documents all stages in the life cycle of a data collection:

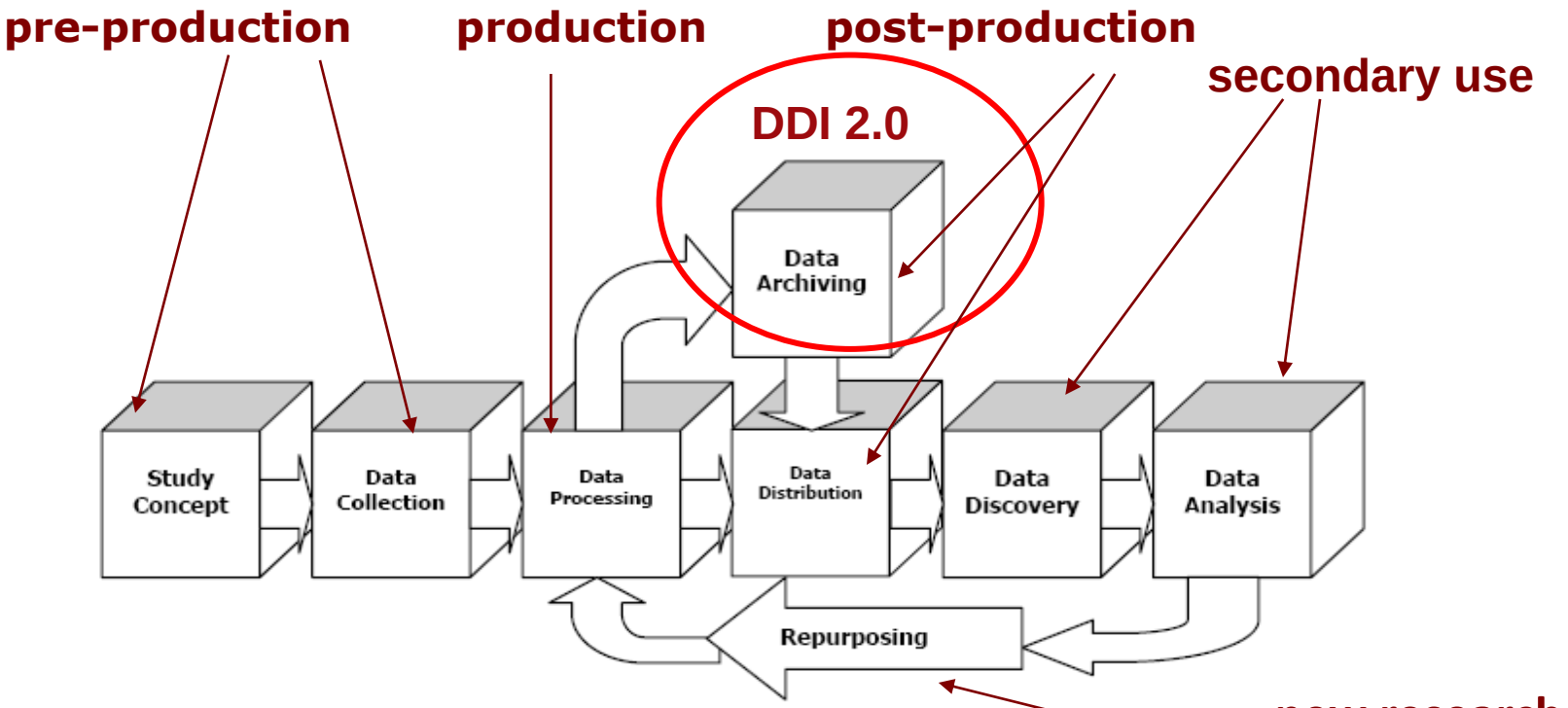
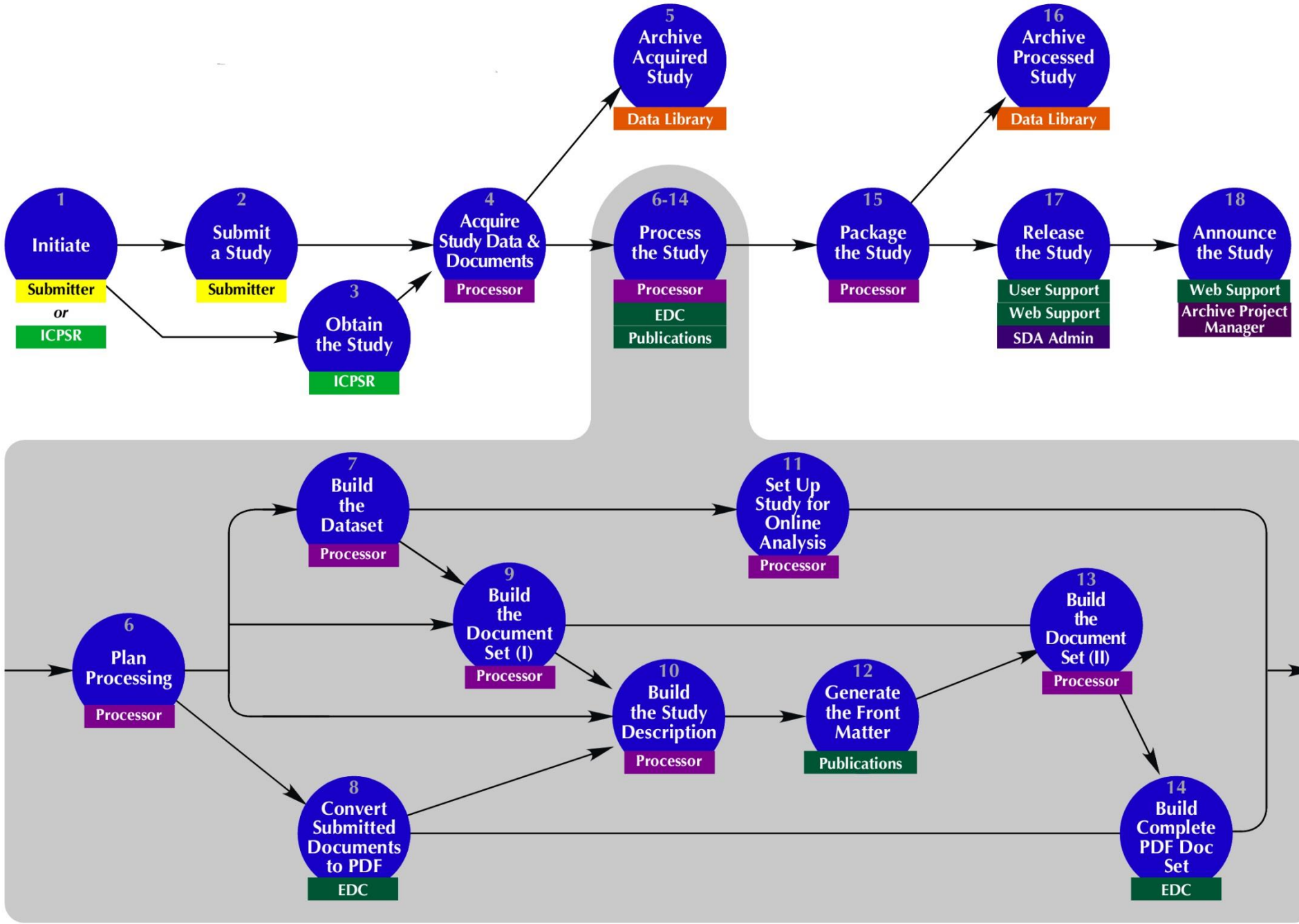



*Figure: Combined Life Cycle Model*




Cole Whiteman • colew@umich.edu • 03/26/2004

# Kontrolni seznam ravnanja s podatki

*Ste pridobili pismeno privolitev?*

*Ste prepričani o tem kdo je lastnik vaših podatkov?*

*Uporabljate standardizirane in konsistentne merske enote in postopke za pridobivanje in preverjanje podatkov?*

*Ste sekundarnim uporabnikom priskrbeli zadostne informacije o metodologiji raziskave, pridobivanju in obdelavi podatkov?*

*So podatki in spremljajoče datoteke označene ustrezno?*

*Ali uporabljate priporočene in sodobne datotečne formate za shranjevanje podatkov?*

*Ali je potrebno vaše podatke anonimizirati?*

*Ali so kopije podatkov – digitalne in fizične – shranjene na varni lokaciji?*

*Ste naredili varnostno kopijo datotek?*

*Ali veste katera različica datoteke s podatki je originalna?*

*Ste vključili tudi informacije za analizo, npr. izpeljane spremenljivke ali povzetke intervjujev?*

# Reference in predlogi za branje


- van den Eynden, V., Corti, L., Woollard, M., Bishop, L. and Horton, L. (2011). Managing Research Data:. Best Practice for Researchers. Colchester: UK Data Archive, University of Essex. [www.data-archive.ac.uk/media/2894/managingsharing.pdf]
- Borgman, Christine L. (2010). Research Data: Who will share what, with whome, when, and why? China-North Americal Library Conference. http://works.bepress.com/borgman/238/]
- National Science Foundation (2010). Data Management & Sharing FAQ. NSF. [http://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp]
- Australian Government, National Health and Medical Research Council, Australian Reseach Council (2007): Australian Code for the Responsible Conduct of Research [http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/r39.pdf]
- Digital Curation Centre – DCC(2010). Data management plans. [http://www.dcc.ac.uk/resources/data-management-plans]
- Jones, Sarah (2011). Develop a Data Management and Sharing Plan. Digital Curation Centre. [http://www.dcc.ac.uk/resources/how-guides/develop-data-plan]
- MIT Libraries (2010). Data management and publishing: organizing your files. [http://libraries.mit.edu/guides/subjects/data-management/organizing.html]