D8 – Report on data collection and preparation: Croatia

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

FNSNF
Swiss National Science Foundation

Deliverable Lead:	UL (ADP), IES
Related Work package:	WP2
Author(s):	Marijana Glavica (FFZG) Alen Vodopijevec (FFZG)
Dissemination level:	Public (PU)
Submission date:	28th April 2017
Project Acronym:	SEEDS
Website:	http://www.seedsproject.ch
Call:	Scientific cooperation between Eastern Europe and Switzerland (SCOPES 2013-2016)
Start date of project:	1st May 2015
Duration:	24 months

Version History

Version	Date	Changes	Modified by
1.0	18.04.2017	Released version	FFZG

Acknowledgments

This report has been developed within the "South-Eastern European Data Services" (SEEDS) (www.seedsproject.ch) project. The participant organisations of the SEEDS project are:

Name	Short Name	Country
Centre for Monitoring and Research, Podgorica	CeMI	Montenegro
Centre for Political Courage, Pristina	CPC	Kosovo
Faculty of Humanities and Social Sciences University of Zagreb	FFZG	Croatia
Institute for Democracy and Mediation, Tirana	IDM	Albania
Institute of Economic Sciences, Belgrade	IES	Serbia
Saints Cyril and Methodius University, Institute for Sociological, Political and Juridical Research, Skopje	ISPJR	Macedonia
Swiss Foundation for Research in Social Sciences, Lausanne	FORS	Switzerland
University of Ljubljana, Social Science Data Archive, Ljubljana	UL	Slovenia

Table of Contents

1. Introduction and history

The first initiative towards the establishment of a data service for social sciences in Croatia started from the researchers themselves 10 years ago. They wanted to have a safe and trusted place where they could share their data and use other's data. Building on that potential, we started to collect the data from these researchers, early supporters, during the SERSCIDA project (2012-2014)[1]. The main focus of this project was to develop processes, tools and an environment for data services in social sciences, to build a prototype data service (website and data archiving system), and to go through an exercise in data curation with one dataset. The SEEDS project for Croatia was a great way to continue efforts towards establishing a data service, and to keep up the promise to early supporters who wanted their data to be published and well documented. We set the goal to manage at least 10 datasets and to make them publicly available. By doing that, we hope to attract more researchers to share and reuse data, but we also want to convince our potential funders about the benefits of data services. To become a trusted service is not possible without the support from local funding bodies, ideally on a national level.

2. Data collection process

During the SERSCIDA project, a few datasets were collected from the researchers who were the initiators of establishing a data service in Croatia. Adding to what we already had, cooperation was starting during the SEEDS project with the founders of the Centre for empirical research in political science (CEPIS)[2] recently established at the Faculty of Political Science, University of Zagreb. An agreement has been reached with CEPIS to make the case for data archiving with the data they planned to make available on their website. Our data archive would apply standard data curation practices to make sure that the data are stored and made available in a secure environment, well documented, and well prepared for long term preservation. CEPIS would publish information (metadata) about the datasets on their website and link to the data archive for data and documentation download. Cooperation with CEPIS have great potential for spreading the culture of data sharing because of their strong empirical orientation and also because some respected journals in the field of political science now require data to be submitted together with a research article.

Later in the project, when the prototype website was created[3] and data archiving tools were ready, an invitation to share data was sent to a broader community of researchers in the area of social sciences. This invitation did not lead to collect much more datasets (just one), but it was a successful reachout in rising awareness about data sharing culture and possibilities. Positive reactions from researchers were received and some promised to look into their data and see what could be shared. Since there are no incentives from policy makers and research funders in Croatia for researchers who share the data, the expectation about collecting more datasets after the first call was not very high. Also, not all researchers are aware of the benefits of data practices, so they need more information before they decide to share their data. That is why within the SEEDS project an individual data

[1] *SERSCIDA project web page http://www.serscida.eu*
[2] *"CEPIS gathers researchers in the field of political science and related disciplines with an interest in designing and implementing empirical research in accordance with the highest international standards. CEPIS is dedicated to analysing democratic political processes and promoting knowledge-based political decision-making in order to raise the quality of public governance in Croatia." http://cepis.hr/en/about-us/*
[3] *SEEDS project WP4 - D11: Report on individual country websites*

service website was to build together with an example collection of datasets as a better way to attract new researchers who are willing to share their data and to use data produced by other researchers.

2.1 Main issues in data collection

Even in the situation where there was a strong initial support from researchers who wanted to share their data, the data collection process was slow. There was a big lesson to learn about how to communicate with researchers and how much time to plan for this process in the future. Researchers who wanted to share their data are usually overly active and they do not have time to go back to previously collected datasets and to properly document them, or to communicate extensively with a data archive about issues in order to provide detailed information about their datasets. In the case of older datasets, they simply do not remember some crucial matters and facts.

Another issue to deal with was related to ethical practices in scientific research. Anonymity has to be guaranteed to participants in social science research. When conducting survey research, personal data is often not even collected, but still it has to be explained to the participants that the survey is anonymous, e.g. that it will not be possible to identify their identity from the answers they give in a survey. A common practice is to stress this out in an introduction part of a questionnaire, where the basic information about the research is given and instructions to participate are also written. One more relevant part of the introduction is the explanation of the purpose of the data collection and how the data will be used. Participants are usually promised that the collected data will be used only for scientific research purposes. This has an impact on defining terms and conditions of use for a particular dataset.

2.2 Defining terms and conditions of use

Datasets and documentation stored in our data repository will be available free of charge. Some limitations and controlled access may be applied, depending on the sensitivity of the data and specific conditions determined by the depositor. The latter are set out in the end user license a version of which all users must accept before being given access to any restricted dataset.

There are three basic levels of access available:

- Research data available without any restrictions (open access)
- Research data available only to registered users
- Research data available only to registered users with special permissions given by the Archive staff or depositor

An access licence has to be defined for each dataset and negotiated with the dataset owners (authors or institutions). If the participants were promised that the data will be used for scientific purposes only, a CCO (Public domain) licence cannot be applied.

3. Collected datasets

We collected 21 datasets, 9 of which are part of the Croatian election studies series, to be included in our public repository. The collected datasets are in the area of sociology and political science, created during various projects led by researchers from academic institutions, or from civil sector

non-profit organisations. A variety of types of data are represented: survey data (quantitative data from online and face to face questionnaires), focus groups transcripts (qualitative data), and a few databases created from publicly available data sources. Most of the datasets were created before the year 2010, and only 5 studies were conducted later. Almost all of the collected survey data (15 out of 16 datasets) were conducted as a national sample, although in 2 cases the sample was not representative.

Within the course of the SEEDS project, we started to manage all collected datasets for archiving in our repository, and 10 of them will be ready for publishing at the end of the project. For the rest of the datasets some additional information from researchers has to be collected and a formal approval from the rights owner has to be obtained before publishing.

The published datasets will be available in our catalogue at the following address: https://dataverse.ffzg.unizg.hr.

4. Data management workflow

The process of receiving datasets from researchers, preparing them for archiving and publishing is one of the crucial parts of a data archiving system. It has to be ensured that all necessary steps in data management are accomplished and that is why a data management workflow was developed. This was done in cooperation with our more experienced partners - ADP and FORS. The starting point was a proposed detailed workflow produced by ADP, based on the SEEDS workshop in Ljubljana[4] together with materials prepared for the RRPP Data Rescue project[5].

There were several versions of the workflow, developed through testing the workflow with real examples, and refined to cover the use of specific tools. The final version of the workflow to be used for managing collected datasets is a list of tasks split into the following categories:

- communication
- general
- materials
- data
- study description
- Nesstar (optional)
- before publishing
- publishing

[4] *D7 - Report on Workshop II_Ljubljana.* http://seedsproject.ch/wp-content/uploads/2015/06/Report-on-Workshop-II_Ljubljana.pdf
[5] *RRPP Data Rescue project.* http://seedsproject.ch/?page_id=618

Communication. The tasks in this category are related to the communication with a data provider. This includes communication about receiving files - inviting researchers to submit their datasets to the archive, discussing how to prepare data (anonymisation and cleaning), ways of transfer, and similar. Another important issue to discuss with the data provider is about access conditions and agreements within the deposit contract.

General. Steps related to storing original documentation and datasets are covered in this category. Also, here we want to make sure that the study is properly identified, e.g. that the basic information for the study is determined (title, year of data collection, author). We may add a study ID and a study acronym manually, or this can be done through the software system. In this category, we also have a separate task for defining data access policies and this is linked to communication with a data provider where the access rights are discussed.

Materials. Here we deal with the preparation of materials (documentation) for ingest. All materials have to be reviewed and it has to be decided which ones are going to be archived. Documents have to be stored in formats appropriate for preservation (AIP) (see D6-Report on integration of technical system).

Data. Data ingest is in a separate category because the procedures for data ingest are different from the one of documentation and the work can be done by different staff members. It has to be checked if anonymisation is done properly and if the submitted files are clean and properly labelled. Originally submitted file formats (usually SPSS, Stata and Excel) have to be converted to the formats acceptable for AIP (see D6).

Study description. This part is about creating metadata for the study description. Minimal and recommended set of fields have to be defined and metadata has to be recorded in a way which enables metadata export in a standard DDI xml format. Metadata will be created in Dataverse.

Nesstar (optional). Different tools can be used for creating study description in XML format and describing variables for a dataset. Using Nesstar Publisher for metadata and ingest procedures is optional and it was not used in our work. The Nesstar workflow can be used in the future to connect our catalogue with the CESSDA Products and Services Catalogue[6] if the Nesstar tools will be somehow used to build it.

Before publishing. In this part of the process we have to check if all permissions for data files are properly set, in accordance to what was agreed with the rights owner. Also, the study description has to be sent to the author for authorisation.

Publishing. After everything has been checked thoroughly, data can finally be published. In this phase, reporting also needs to be done, and preservation of all materials. The report has to be sent to the researcher together with clean files.

[6] *CESSDA Data Catalogue : https://cessda.net/CESSDA-Services/Resources/Data-Catalogue*

5. Tools

5.1 Workflow implementation in Redmine

Redmine is an open source project management tool[7]. It was used to support our data management workflow and to guide us through the process. It allows us to record and document all work that has been done around each dataset, and also to record our progress in managing datasets by which it reminds us what remains to be done. Redmine was also used for testing the data management workflow which can still be further improved.

Each dataset is managed in a separate project in Redmine, except for a series of election studies which was covered as one project. Groups of tasks in our workflow were represented as trackers. In each tracker, several issues were created to describe tasks which have to be accomplished for that part of the workflow. Issues can be hierarchically organised, so child issues (subtasks) can be created for each issue when needed. Also, issues can have related issues in which way we can easily see connected or dependent tasks.

Short specific instructions were developed for most of the tasks, except for Nesstar, because we didn't use it for now. The instructions were also revised during the development of the workflow.

The workflow, as implemented in Redmine, allows delegating tasks to different staff members. Each task has one assignee, which is responsible for getting the task done and recording work and progress.

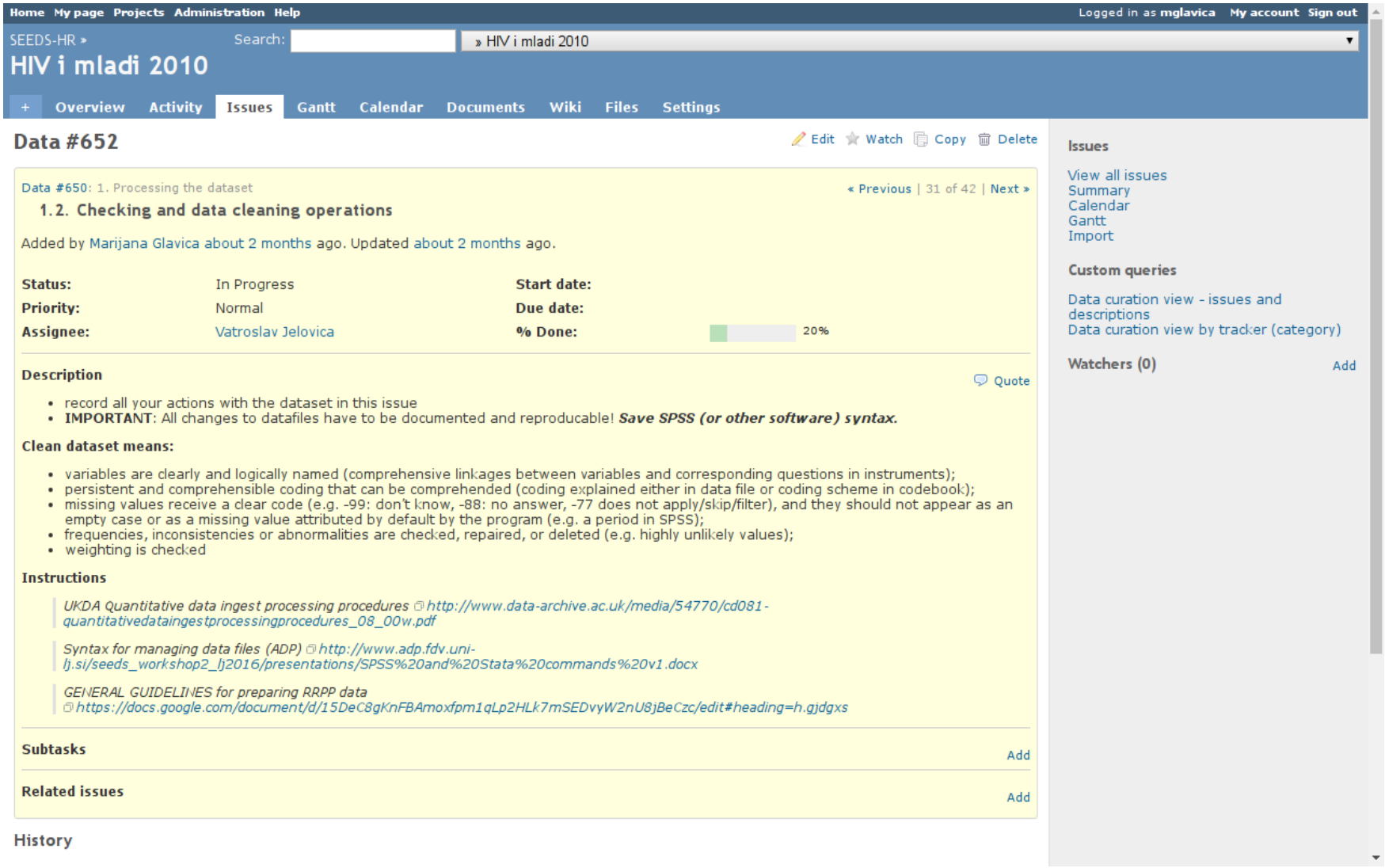
An example issue in Redmine is shown in Picture 1. First we see the status of the task, which is the assignee for the task, and what progress has been made. A short description and instructions for the task are below, followed by the part where the work is recorded, and questions are asked and discussed.

5.2 Repository building in Dataverse

Dataverse is an open source web application to share, preserve, cite, explore, and analyse research data.[8] It was selected during the evaluation process (see D9) as an archiving and dissemination tool for the future data archive. How the tool intended to be used by the Croatian data archive is explained in D6.

Parallel to the SEEDS project activities, cooperation has been established with DANS - Data Archiving and Network Services from Netherlands who started the Dataverse pilot project as an activity of the CESSDA SaW project. Through this pilot project, DANS was exploring possibilities to become a support centre for Dataverse and our Dataverse installation (together with Serbian) served as an example implementation for European data archives.

[7] *Redmine - project management web application.* http://www.redmine.org/
[8] http://dataverse.org/about



Picture 1: Example task (issue) in Redmine

Organising dataverses and datasets

In the Dataverse software, datasets are organised in so called dataverses. A dataverse is a container for datasets and other dataverses[9] and it can be of a different type: institution or organisation, laboratory, research group, research project, researcher, or journal. One of the first steps, right after the installation and system configuration in the Dataverse implementation, was to define the organisational structure of dataverses which contain datasets and other dataverses. Our repository aims to become a national service, but also we want to enable institutions to build and administer their own dataverses and datasets (in other words, to build their own institutional repositories). A similar approach was taken by another important infrastructure project in Croatia - Dabar[10], which currently offers infrastructure for building institutional repositories of publications but also aims to include some kind of support for datasets. In order to have a structure which is compatible with the structure of the Dabar repository, to facilitate exchange of data and metadata, datasets were organised into dataverses which represent the affiliation with the main author (principal investigator). A hierarchical organisation allows going deeper into smaller units of the organisation.

Study description

Dataverse supports metadata creation for the study and it is able to export metadata to a standard DDI xml file. A set of domain independent metadata fields is available for use by default and this can be configured to define required and optional fields. In addition to that, a basic set of metadata fields specific for social sciences and humanities research is available and it was used for our purposes.

Uploading files

Any file format can be uploaded to Dataverse. A data service should have a list of accepted formats. For our future data services this can be found in the D6 - Report on integration of technical system.

Tabular data files ingest

Tabular data files (Stata, SPSS, R, Excel/xlsx and CSV) are supported in Dataverse by an additional functionality. During the ingest phase of this file formats, files are processed and data content is extracted from other descriptive information (descriptive labels, categorical values and labels, and more) recorded in original format files. Data is converted into tabular data which is an archival format suitable for long time preservation. Other information describing the content of the data file is stored separately as metadata which can be exported as plain text XML files in standard DDI Codebook format.

Storing data and metadata in non-proprietary formats and separating them to data and metadata also enables the use of additional tools for data exploration. Dataverse is integrated with TwoRavens Data Exploration and Analysis Tool[11] and we are planning to install it for our future data service.

[9] *Dataverse User Guide: Dataverse Management : http://guides.dataverse.org/en/4.6.1/user/dataverse-management.html*
[10] *DABAR (Digital Academic Archives and Repositories) is the key component of the Croatian e-infrastructure's data layer. It provides technological solutions that facilitate maintenance of higher education and science institutions' digital assets, i.e., various digital objects produced by the institutions and their employees: http://dabar.srce.hr*
[11] *TwoRavens Data Exploration Tool: http://2ra.vn/*

Terms of use

Based on what was agreed with data owners, terms of use have to be outlined for each dataset which explain how the data can be used once downloaded. Dataverse default is CC0 public domain dedication licence which facilitates reuse and extensibility of research data, but it is not always possible to publish a dataset under such conditions. In that case, a custom licence can be entered.

For each file in a dataset, terms of access can also be set, providing information on how and if users can gain access to restricted files.

Managing permissions

Permissions in Dataverse can be set up for dataverses and for datasets. To control access, different roles can be assigned to users and/or groups. On a dataset level, permission can be set up for the whole dataset or for a particular file.

Persistent identifiers (DOIs)

Persistent identifiers are an integral part of the Dataverse system. This means that without a registered DOI namespace it is not possible to publish a dataset. DOIs for the objects in our data repository will be registered through the da|ra[12] registration agency.

[12] *da|ra - registration agency for social and economic data: https://www.da-ra.de/en/*

Appendix I. List of datasets prepared for publishing

Dataset No. 1
Title: Seksualna uporaba interneta, 2004 [Sexual use of the Internet, 2004]
Principal investigator: Aleksandar Štulhofer
Institution: Faculty of Humanities and Social Sciences, University of Zagreb
Description: The first Croatian study of online sexual activities (OSA). The sample consisted of 2079 female and male members of the most popular Croatian dating website. Age and gender structure of the sample closely resembled the national population of Internet users.
Type of data: survey data
Access rights: Open Access

Dataset No. 2
Title: HIV i mladi 2005 [HIV and youth 2005]
Principal investigator: Aleksandar Štulhofer
Institution: Faculty of Humanities and Social Sciences, University of Zagreb
Description: The research was conducted on youth population aged 18 to 24 years. Sexual behaviour, knowledge and attitudes of Croatian youth were examined.
Type of data: survey data
Access rights: For research purposes only

Dataset No. 3
Title: Anketa o obrazovnim i radnim karijerama mladih u Hrvatskoj, 2008 [Survey of the educational and work careers of young people in Croatia, 2008]
Principal investigator: Teo Matković
Institution: Faculty of Law, University of Zagreb
Description: The target population of this survey wes those who left education in the last five years regardless of whether the last level of education attended was completed successfully or unsuccessfully.
Type of data: survey data
Access rights: For research purposes only

Dataset No. 4
Title: HIV i mladi 2010 [HIV and youth 2010]
Principal investigator: Aleksandar Štulhofer
Institution: Faculty of Humanities and Social Sciences, University of Zagreb
Description: The main objective of the study was to repeat the survey among the youth population in the Republic of Croatia that was conducted in 2005 and to determine the level of knowledge about AIDS, the main features of sexual behaviour among young people in order to improve HIV prevention in the country.
Type of data: survey data
Access rights: For research purposes only

Dataset No. 5
Title: Elektroničko glasovanje zastupnika i zastupnica 6. saziva Hrvatskog sabora (2008-2011) [Electronic voting of representatives in 6th Croatian Parliament (2008-2011)]
Principal investigator: Daniela Širinić
Institution: Faculty of Political Science, University of Zagreb
Description: The database encompasses the data on the voting of individual MPs for the 6th convocation of the Croatian Parliament (2008-2011) from the 4th to 24th Session of the Parliament.
Type of data: The data was extracted from the electronic voting register in cooperation with the Information Documentation Service of the Parliament.
Access rights: Open Access

Dataset No. 6
Title: Baza podataka o parlamentarnoj političkoj eliti u Hrvatskoj od 1990. do 2015. [Database about parliamentary political elite in Croatia from 1990 to 2015]
Principal investigators: Goran Čular (Fakultet političkih znanosti), Vlasta Ilišin (Institut društvenih znanosti Zagreb), Višeslav Raos (Fakultet političkih znanosti)
Institution: Faculty of Political Science, University of Zagreb
Description: The database on the parliamentary political elite in Croatia contains detailed information on the MPs of all convocations of the Parliament from 1990 to 2015.
Type of data: biographic and demographic information

Access rights: Open Access

Dataset No. 7
Title: Izbori za predsjednika političkih stranaka u Hrvatskoj 1989-2016 [Elections of the President of political parties in Croatia from 1989 to 2016]
Principal investigator: Dario Nikić Čakar
Institution: Faculty of Political Science, University of Zagreb
Description: The database includes seven relevant parties in Croatia that have organizational continuity throughout the observed period and who regularly competed in the elections and won mandates. Seven parties (HDZ, SDP, HSS, HSP, HSLS, IDS and HNS) and 66 presidential candidates have been included.
Type of data: voting data - source of data are the minutes of the election assemblies of the political parties and on the reports on the elections that are available in the archives of the political parties or are submitted to the Ministry of Public Administration of the Republic of Croatia.
Access rights: Open Access

Dataset No. 8
Title: Odnos građana RH prema EU i pristupanju/članstvu Hrvatske EU [Attitudes of Croatian citizens towards the EU and Croatian membership in EU]
Principal investigator: Benjamin Čulig
Institution: Faculty of Humanities and Social Sciences, University of Zagreb
Description: National study of information, attitudes and beliefs of Croatian citizens related to the European Union, Croatia's accession to the European Union and Croatia's membership in European Union.
Type of data: survey data
Access rights: For research purposes only

Dataset No. 9
Title: Sharing and Use of Data in Social Sciences - Researchers' Practices and Needs in Bosnia and Herzegovina, Croatia and Serbia, 2012
Investigators: SERSCIDA project team
Description: Survey on production, preservation and use of research data in the field of social

sciences. Information about the researchers' experience of documenting, re-use and disseminating of research data, their use of statistical/analytical software packages, preferred methodology, and their institution's policies regarding long term preservation and/or documentation of data.
Type of data: survey data
Access rights: Open Access

Dataset No. 10
Title: Anketno istraživanje o radnim karijerama diplomiranih studenata, 2015 [Survey on Working Careers of Graduate Students, 2015]
Principal investigator: Dragan Bagić
Institution: Faculty of Humanities and Social Sciences, University of Zagreb
Description: In this survey, information on working careers and student occupations was gathered. The survey included the population of students who graduated from selected study groups between 2003 and 2014, regardless of the year of study enrolment.
Type of data: survey data
Access rights: For research purposes only