

# Warsztat badacza: Repozytoria danych

22 maja / 16 czerwca 2020 r.

Marta Kołczyńska

Instytut Studiów Politycznych Polskiej Akademii Nauk

[mkolczynska@gmail.com](mailto:mkolczynska@gmail.com)



# Plan dnia

Dane: Bardzo krótkie wprowadzenie

- Typy danych
- Dane – metadane – paradane – dokumentacja
- Jakość danych

Wyszukiwanie danych

Repozytoria danych sondażowych

Zadanie „domowe”

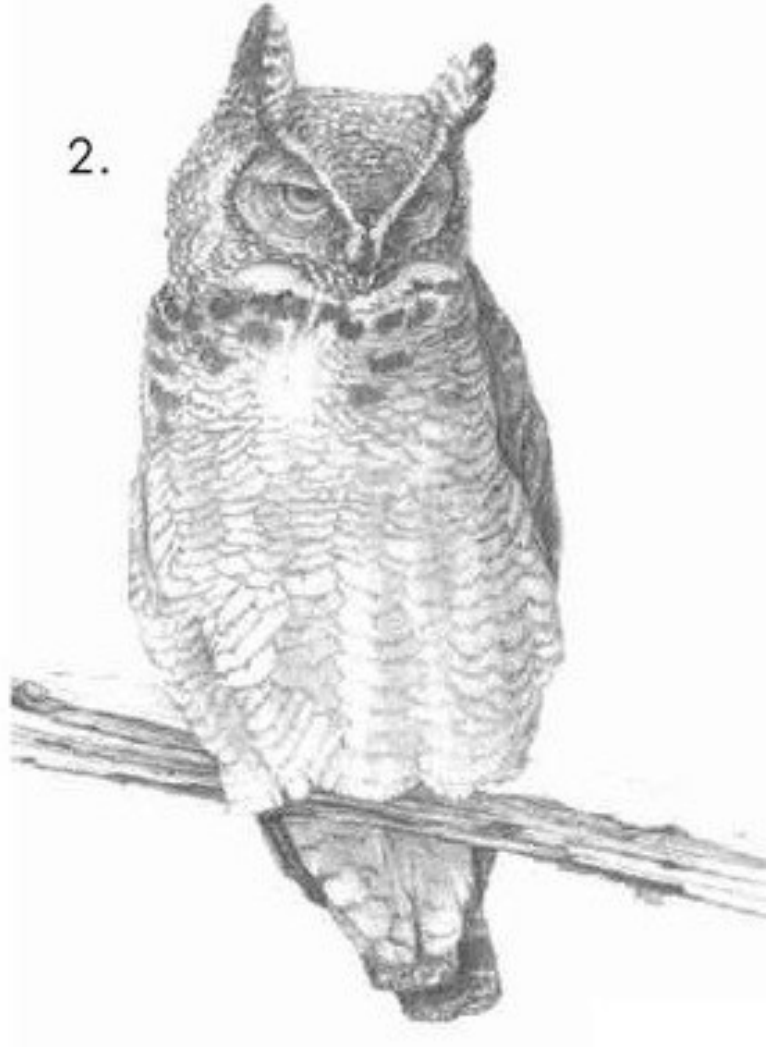
## How to draw an owl

1.



1. Draw some circles

2.



2. Draw the rest of the owl

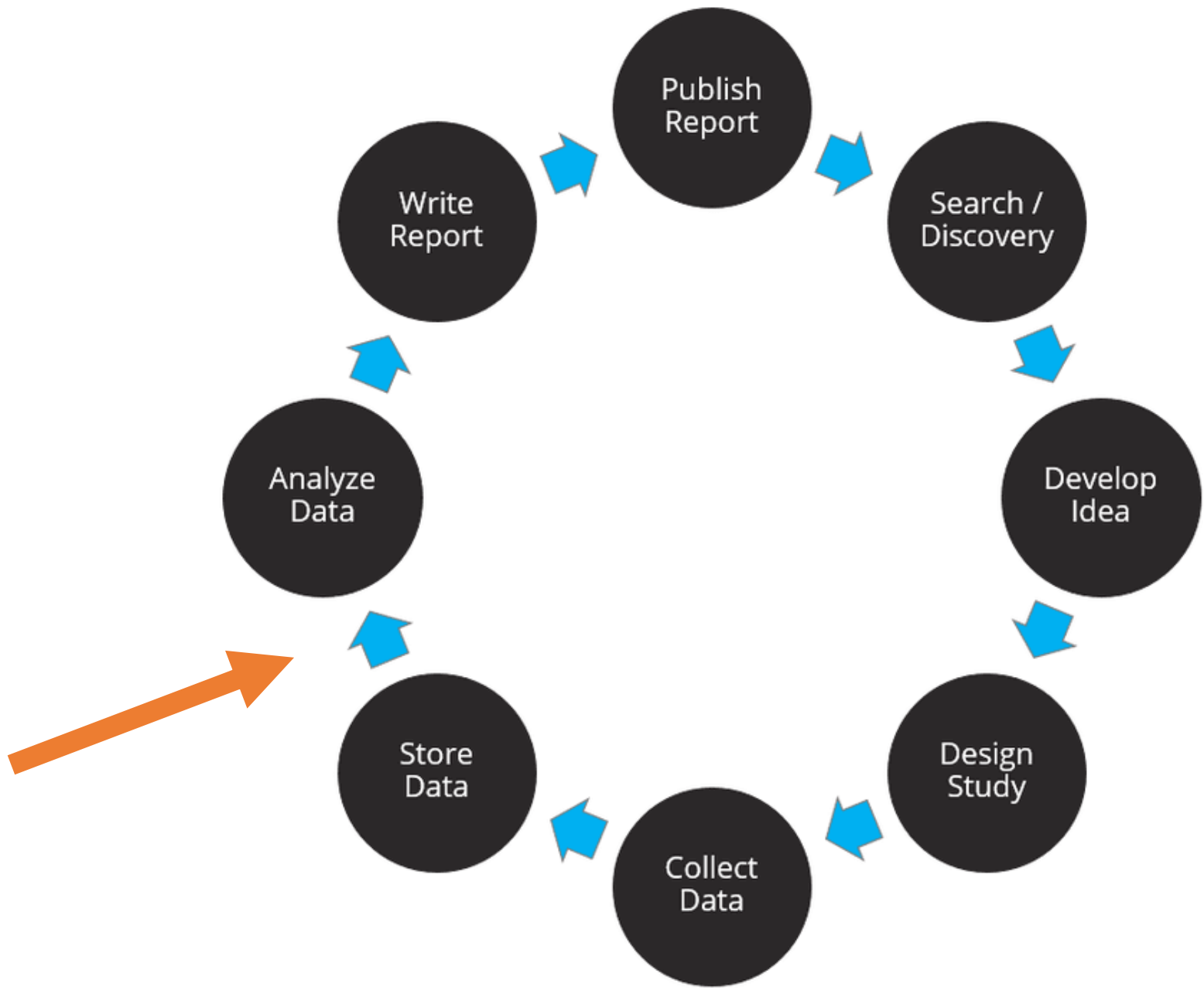
# Typy danych / analiz

## Pierwotne

- Dane zebrane i analizowane przez tę samą osobę/zespół
- Badanie zaprojektowane w konkretnym celu

## Wtórne

- Analiza danych zastanych, zebranych i zarchiwizowanych przez inną osobę/zespół



# Zalety analiz wtórnych

- Recycling: oszczędność czasu i pieniędzy
- Ułatwienie analiz porównawczych (np. międzykrajowych) i podłużnych/historycznych
- Możliwość wzbogacenia badań o analizę danych różnego typu (np. ilościowych i jakościowych)
- Minimalizuje obciążenie respondentów

# Wady analiz wtórnych

- Dane zostały zebrane w innym celu i są nieoptymalne
- Konieczność wyczyszczenia/uzdatnienia/uzupełnienia danych, czasem przy ograniczonej dokumentacji
- Brak kontroli nad jakością danych i ograniczona możliwość jej oceny
- Konieczność interpretacji danych biorącej pod uwagę kontekst ich powstawania
  - Dane mogą być ograniczone czy obciążone przez decyzje polityczne minionych epok, np. pomijając pewne grupy ludności
  - Definicje mogą podlegać zmianie (np. definicje przestępstw)

# Analizy w oparciu o dane wtórne

Praktycznie wszystkie typy analiz, ale ograniczeniem jest dostęp do danych różnych typów.

Większość danych wtórnych to dane ilościowe.

Dane z badań etnograficznych, np. obserwacji uczestniczących, raczej nie nadają się do analizy wtórnej, chociaż mogą być ciekawym materiałem pomocniczym.

Kwestia bezpieczeństwa i zapewnienia anonimowości.

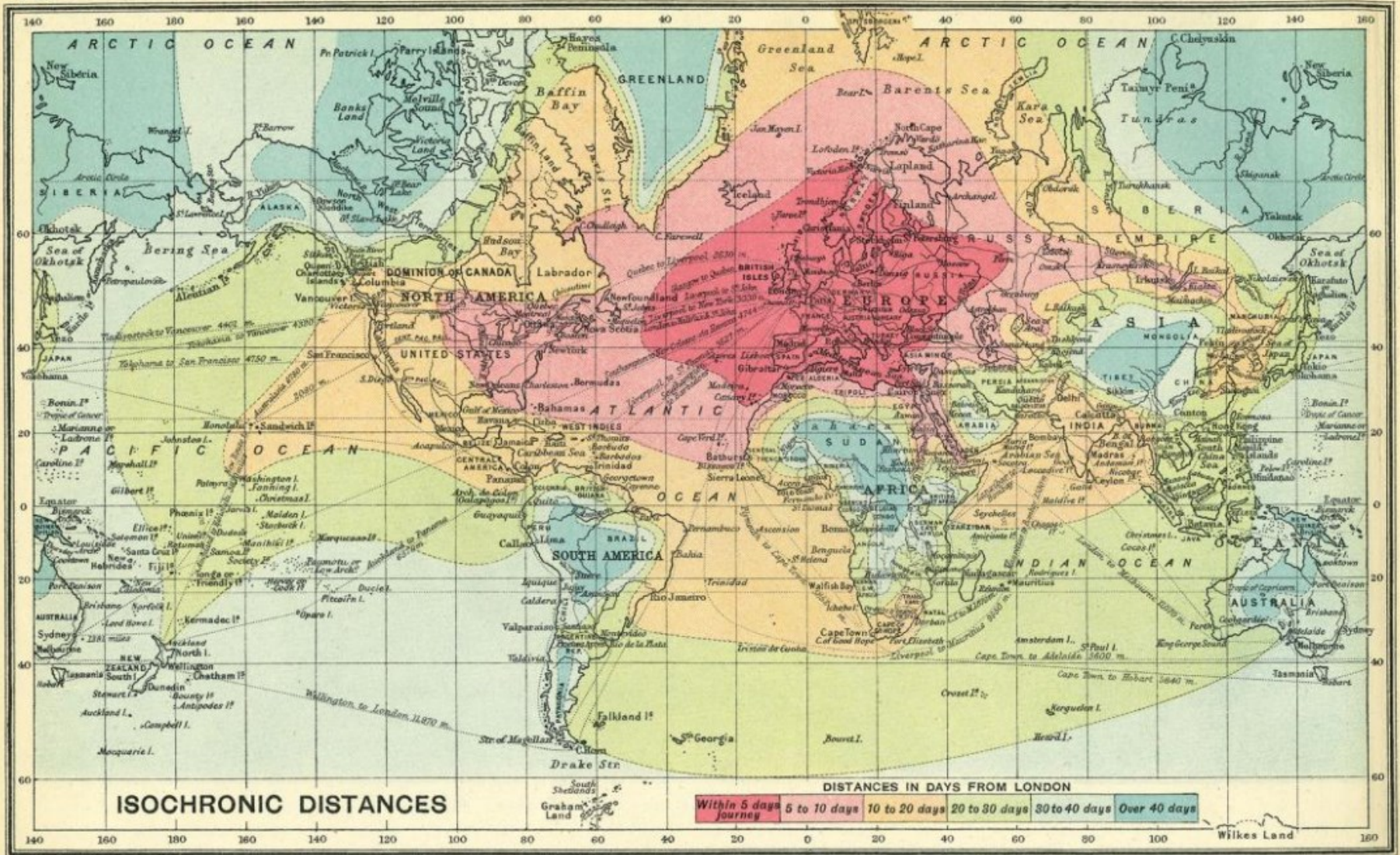


# Dane wtórne w socjologii

- Dane sondażowe
  - Dane wielokrajowe
  - Dane panelowe
  - Badania wykorzystywania czasu (budżet czasu ludności)
- Dane opisujące kraje, w tym dane administracyjne
- Dane jakościowe: wywiady, pamiętniki
- Również: zdjęcia, nagrania audio i wideo, wszelkie inne materiały stanowiące podstawę badań
- Publikowane wyniki badań (→ metaanaliza)

# Czas podróży w dniach z Londynu w 1914

12<sup>B</sup>



John G. Bartholomew. 1914. "An Atlas of Economic Geography", za Simon Willis. 2015. Time Travel.

<https://www.1843magazine.com/places/cartophilia/time-travel>

# Czym są dane?

Dane = fizyczna reprezentacja danych w formie umożliwiającej komunikację oraz przetwarzanie

(UNECE 2000: 6)

Dane badawcze = dane wykorzystywane do realizacji badań naukowych oraz oceny ich wyników

Dane

Metadane = dane o danych

Dokumentacja

## Czym są metadane?

*“Metadane to ustrukturyzowane informacje opisujące, tłumaczące, lokalizujące i ułatwiające we wszelki inny sposób odnalezienie, wykorzystanie lub zarządzanie zasobem informacji. Metadane często określa się mianem ,danych o danych’ albo ,informacji o informacjach’.”*

*-- National Information Standards Organization*

<http://www.niso.org/publications/press/UnderstandingMetadata.pdf>

Metadane dostarczają informacji umożliwiających uporządkowanie **danych** (np. dokumentów, plików graficznych, zbiorów danych), **pojęć** (np. schematów klasyfikacyjnych) i **elementów świata rzeczywistego** (np. ludzi, organizacji, miejsc, obrazów, produktów).

# Przykłady metadanych

Etykieta



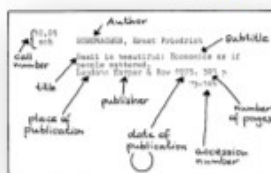
dostarcza metadanych  
na temat



puszki



Karta katalogowa



książki



Opis zbioru danych (DCAT)

```
:weather1-7 a dcat:Dataset ;
dct:title "Measurements from weather stations 1-7" ;
dct:description "Data from seven weather stations
showing temperature, humidity,
wind direction and wind speed" ;
dct:modified "2013-07-01" ;
dct:publisher <http://myweather.com/id/myweather> ;
dcat:keyword "weather" ;
dcat:landingpage <http://myweather.com/stations1-7.html> ;
dcat:distribution :weatherdata-xlsx
.

:weatherdata1-7-xlsx a dcat:Distribution ;
dct:format <http://publications.europa.eu/resource/authority/file-type/XLSX> ;
dct:licence <http://creativecommons.org/licenses/CC0> ;
dcat:downloadURL <http://myweather.com/stations1-7.xlsx>
.
```



zbioru danych

	Temp. °C	Humidity %	Wind direction	Wind speed km/h
Station 1	18.1	60	WSW	18
Station 2	17.5	59	WSW	20
Station 3	18.2	55	SW	22
Station 4	19.9	62	TW	18
Station 5	18.8	65	WSW	19
Station 6	18.2	63	SSW	23
Station 7	17.9	61	SW	22

# Metadane danych badawczych

## Opis badania

- Cel, producent, finansowanie, zakres, czas
- Populacja, dobór próby, jednostka obserwacji, sposób zbierania danych

## Informacja o narzędziach badawczych

- Kwestionariusze, karty odpowiedzi, instrukcje dla ankieterów

## Opis plików danych

## Opis zmiennych

- W danych: nazwy zmiennych, etykiety zmiennych, etykiety wartości
- Dokumentacja: słownik / codebook, kody braków danych
- Informacja o wprowadzaniu danych, korygowaniu błędów

Paradane (dane o procesie zbierania danych), informacje kontekstowe

Informacje o możliwości wykorzystywania i przetwarzania

# Jakość danych

Dokładność (accuracy)

Spójność (consistency)

Dostępność (availability)

Zupełność (completeness)

Zgodność ze standardami (conformance)

Wiarygodność (credibility)

Przetwarzalność (processability)

Odpowiedniość/adekwatność (relevance)

Aktualność (timeliness)

Wyszukiwanie danych





Leaders

Feb 25th 2010 edition >

Technology

# The data deluge

# Producenci danych

Instytucje i organizacje, w ramach swojej działalności

Dane wytwarzane jako część infrastruktury badawczej

Dane wytwarzane przez projekty badawcze

Dane generowane automatycznie, np. historia edycji Wikipedii

Ruch otwierania danych

# Reakcje na nadmiar

## Standardy

- DDI = Data Documentation Initiative

## Wyszukiwarki

## Trwałe identyfikatory

- DOI = Digital Object Identifier

# Wyszukiwanie danych

## Instytucje i organizacje

- Urzędy statystyczne (np. Główny Urząd Statystyczny, Eurostat)
- Organizacje międzynarodowe (np. Bank Światowy, ONZ, WHO)



Narodowy Spis Powszechny Ludności i Mieszkań 2021 • Powszechny Spis Rolny 2020 • Badania statystyczne



Podstawowe dane



Opracowania sygnałne



Publikacje



Bank Danych Lokalnych



Bank Danych Makro-ekonomicznych



SDG



Działynowe Bazy Wiedzy



STRATEG



Portal Geostatystyczny



Portal API



Dashboard gospodarczy



REGON, TERYT

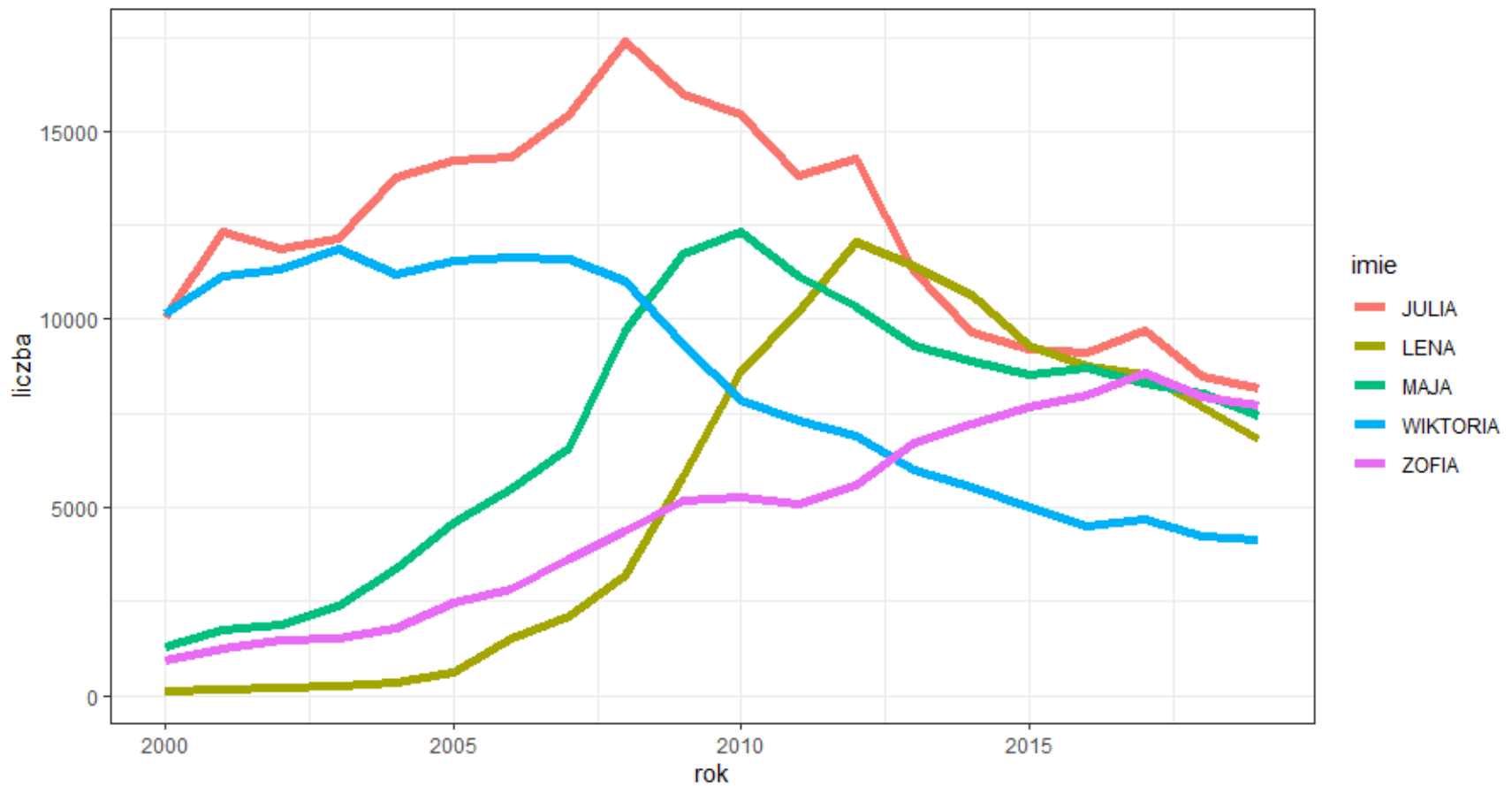
- Otwarte dane, np. [dane.gov.pl](https://dane.gov.pl)

Sort by: Data date ▾

[Imiona męskie nadane dzieciom w Polsce w 2019 r. wg województw - imię drugie](#)Data date: 21 January 2020 Download:  
[XLSX \(73kB\)>>](#)[Imiona żeńskie nadane dzieciom w Polsce w 2019 r. wg województw - imię drugie](#)Data date: 21 January 2020 Download:  
[CSV \(96kB\)>>](#)[Imiona męskie nadane dzieciom w Polsce w 2019 r. - imię drugie](#)Data date: 21 January 2020 Download:  
[XLSX \(21kB\)>>](#)[Imiona męskie nadane dzieciom w Polsce w 2019 r. wg województw - imię drugie](#)Download:  
[CSV \(102kB\)>>](#)

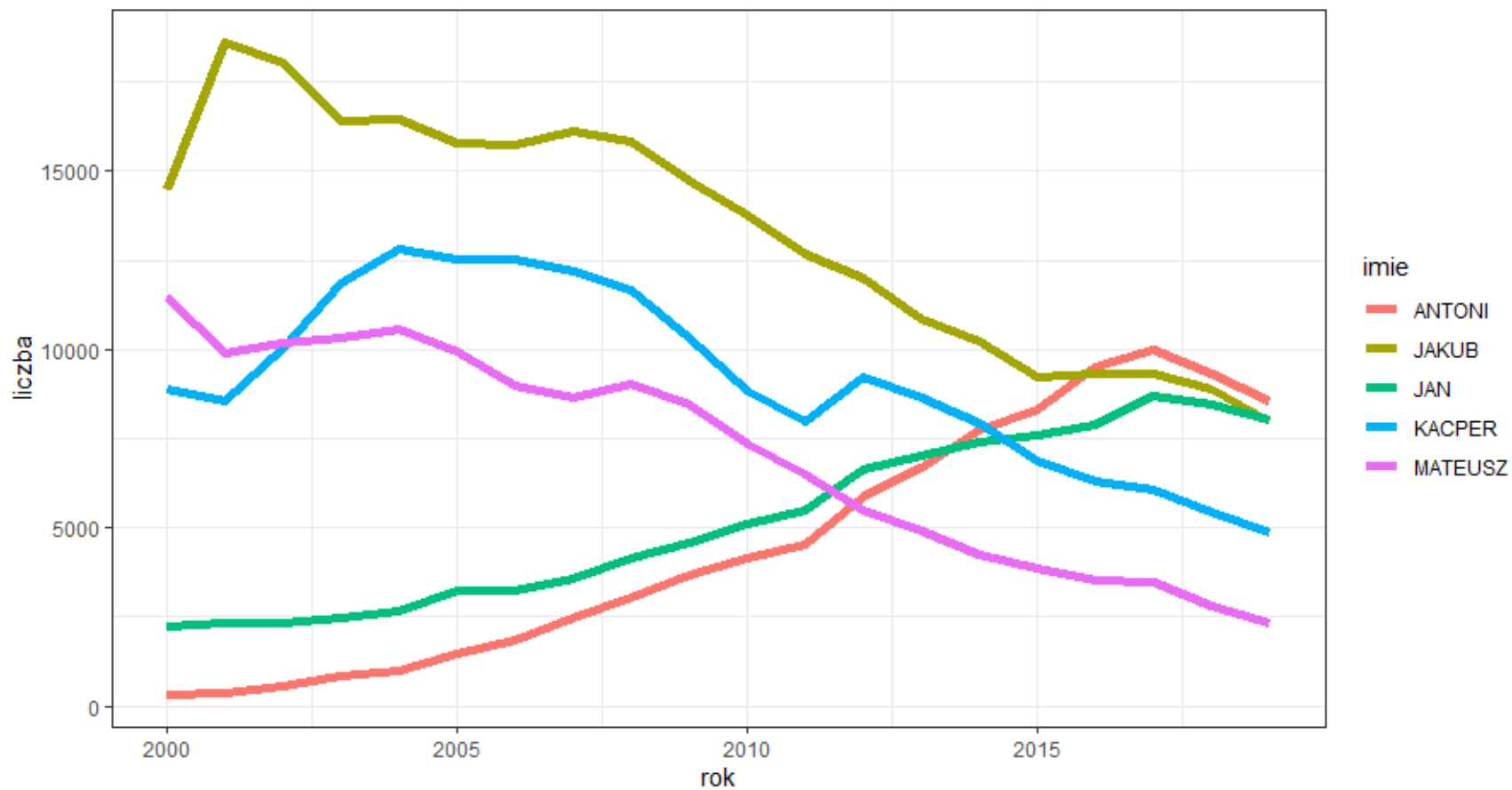
# Imiona nadane dzieciom w Polsce w latach 2000-2019 - imię pierwsze

Imiona nadane dzieciom w Polsce w latach 2000-2019 wg pola: imię pierwsze, z podziałem na płeć.



# Imiona nadane dzieciom w Polsce w latach 2000-2019 - imię pierwsze

Imiona nadane dzieciom w Polsce w latach 2000-2019 wg pola: imię pierwsze, z podziałem na płeć.



# Wyszukiwanie danych

## Archiwa danych

- Materiały edukacyjne, szkolenia, itd.
- Archiwum Danych Społecznych
- Archiwum Danych Jakościowych
- Archiwum Badań nad Życiem Codziennym
- GESIS - Leibniz Institute for the Social Sciences (Niemcy)
- UK Data Archive (Wielka Brytania)
- Inter-university Consortium for Political and Social Research (ICPSR, USA)
- Consortium of European Social Science Data Archives (CESSDA)

Wyszukiwarka repozytoriów danych (Registry of Research Data Repositories):

<https://www.re3data.org/>



# Archiwum Danych Społecznych (UW & IFiS PAN)



## Ostatnio zdeponowane:

- 30.11.2017 **East European Parliamentarian and Candidate Data (EAST PaC) 1985 - 2015**, IFiS PAN.
- 30.11.2017 **Opinie i oczekiwania wobec edukacji dotyczącej rozwoju psychoseksualnego i seksualności**, IBE.
- 25.05.2016 **Zdrowie, nierówności, dezintegracja społeczna (Mieszkańcy Warszawy 2003)**, IFiS PAN.
- 13.05.2016 **East European Parliamentarian and Candidate Data (EAST PaC) 1985 - 2014**, IFiS PAN.
- 12.03.2014 **Zoom na Uniwersytety Trzeciego Wieku**; Towarzystwo Inicjatyw Twórczych "ę"
- 17.01.2014 **Struktura społeczna w Polsce: POLPAN 2008**; Zespół Porównawczych Analiz Nierówności Społecznych, IFiS PAN.
- 23.09.2013 **Polskie Generalne Sondaże Społeczne 1992-2010** (ISS UW)
- 29.01.2013 **Polskie Generalne Studium Wyborcze (PGSW) 2011**, Instytut Studiów Politycznych Polskiej Akademii Nauk.
- 22.01.2013 **International Social Survey Programme 2007** (Social Inequality IV), 2008 (Religion III) i 2009 (Leisure Time and Sports).
- 21.07.2011 **Aktualne Problemy i Wydarzenia** (CBOS) 2008
- 24.05.2011 **„Solidarność” – doświadczenie i pamięć**, (CBOS i ECS) 2010.
- 21.02.2011 **Diagnoza Społeczna 2000-2009. Warunki i Jakość Życia Polaków** (RMS)
- 21.02.2011 **Polskie Generalne Studium Wyborcze (PGSW) 2007**

# Archiwum Danych Społecznych (UW & IFiS PAN)



## Podręcznik archiwizacji danych społecznych

**Bogdan Cichomski, Tomasz Jerzyński, Marcin Zieliński**

Zespół Ośrodka Badań Socjologicznych  
Instytutu Studiów Społecznych  
Uniwersytetu Warszawskiego

Wersja 1

Tekst podręcznika dostępny jest także w formacie dokumentu PDF

## Spis treści

1. Wstęp
  2. Korzyści archiwizacji danych
  3. Znaczenie planowania zarządzaniem danymi
    - 3.1. Dokumentacja jako część planu projektu
    - 3.2. Oprogramowanie
  4. Kontrola poprawności danych w zbiorze i jego spójności
    - 4.1. Kontrola błędów w zbiorze
      - 4.1.1. "Dzikie kody"
      - 4.1.2. "Outliers"
    - 4.2. Kontrola spójności logicznej
      - 4.2.1. Pytania filtrujące
      - 4.2.2. Instrukcje przed pytaniem
      - 4.2.3. Daty i przedziały czasu
    - 4.3. Kontrola zbiorów o strukturach hierarchicznych
  5. Przygotowanie dokumentacji techniczno-metodologicznej badania
    - 5.1. Książka kodów (codebook)
    - 5.2. Rozkłady częstości (frequencies)
    - 5.3. Statystyki opisowe (descriptive statistics)
  6. Przygotowanie zbiorów do archiwizacji
    - 6.1. Akceptowane formaty zbiorów danych
    - 6.2. Techniczne aspekty zbiorów danych
      - 6.2.1. Słownik definiujący dane (data definition statements)
      - 6.2.2. Nazwy zmiennych (variable names)
      - 6.2.3. Etykiety zmiennych (variable labels)
      - 6.2.4. Wartości zmiennych (values)
      - 6.2.5. Etykiety wartości (value labels)
      - 6.2.6. Braki danych i kody specjalne (missing values)
        - 6.2.6.1 Wybór kodów specjalnych
  7. Zarządzanie danymi
  8. Kopie zapasowe
  9. Anonimizacja
  10. Sumaryczne zestawienie wymagań stawianych deponentom
    - 10.1. Zbiór danych
    - 10.2. Dokumentacja
  11. Przekazywanie danych do Archiwum
    - 11.1. Informacje ogólne
    - 11.2. Deponowanie danych zarchiwizowanych na płycie CD
    - 11.3. Przesyłanie danych z wykorzystaniem protokołu FTP
      - 11.3.1. Konfiguracja programu FTP do pracy z serwerem ADS
      - 11.3.2. Logowanie się na serwer ADS i upload zbiorów danych
- Aneks A  
Aneks B



## Archiwum gromadzi:

### 1) Zbiory pochodzące z historycznych, kluczowych dla polskiej socjologii i antropologii badań jakościowych

W ramach projektu "Archiwum Danych Jakościowych przy IFiS PAN", realizowanego w latach 2012-2018, zarchiwizowaliśmy i udostępniliśmy następujące kolekcje:

Wybór materiałów terenowych Józefa Obrębskiego z badań w Macedonii w latach 1932-1933

Materiały terenowe Józefa Obrębskiego z badań na Polesiu w latach 1934-1937

Wybór autobiografii chłopskich z konkursu pamiętnikarskiego ogłoszonego przez Insytut Kultury Wsi w roku 1937

Materiały terenowe Stanisława Ossowskiego z badań w Purdzie Wielkiej na Warmii w latach 1948-1949

Udostępniliśmy również zarchiwizowany wcześniej w ramach innego zrealizowanego przez nasz zespół projektu zbiór „Badania stylu życia prof. Andrzeja Sicińskiego z lat 1976-1980”, będący częścią spuścizny naukowej dawnego Zespołu Badań nad Stylami Życia przy IFiS PAN (badania z lat 70. i 80.).

Osoby zainteresowane informacją na temat dostępności innych materiałów z klasycznych badań antropologicznych i socjologicznych zapraszamy [tutaj](#).

### 2) Zbiory powstałe po 1989 roku, zarówno te pochodzące z końca XX w., jak i całkiem współczesne

Część z nich pochodzi z badań przeprowadzonych przez ważne dla socjologii osoby, które niestety już odeszły, takie jak Antronina Kłoskowska czy Elżbieta Tarkowska. Inne zostały wytworzone przez czynnych wciąż zawodowo uczonych przed upowszechnieniem cyfrowych metod rejestracji materiałów. Ostatnią grupę, która, jak mamy nadzieję, będzie się rozrastać, stanowią materiały z badań jakościowych zakończonych w ostatnich latach. Zapraszamy do przejrzienia katalogu!

[Data Catalogue](#)[Data Management Expert Guide](#)[Training](#)

[Spread of COVID-19 in Austria. PCR-tests in a representative sample \(SUF edition\)](#)

*SORA Institute for Social Research and Consulting (SORA Institute for Social Research and Consulting)*

\* PCR = Polymerase Chain Reaction. Wykrywanie materiału genetycznego.

[GESIS Panel Special Survey on the Coronavirus SARS-CoV-2 Outbreak in Germany](#)

*GESIS Panel Team (GESIS Leibniz-Institut für Sozialwissenschaften)*

[InfraRisk Module B - Calculation of Avalanche and Rockfall Prone Roads and Railways, 2010](#)

*Frauenfelder, Regula (Norwegian Geotechnical Institute); Solheim, Anders (Norwegian Geotechnical Institute)*

**Study title**

Spread of COVID-19 in Austria. PCR-tests in a representative sample (SUF edition)

**Creator**

SORA Institute for Social Research and Consulting (SORA Institute for Social Research and Consulting)

**Study Persistent Identifier**

doi:10.11587/X2MIHW (DOI)

**Abstract**

Full edition for scientific use. The main objective of this study was to give an estimate of the spread of COVID-19 ("Corona Virus") among non-hospitalized people living in Austria. This study is the first in continental Europe based upon nationwide PCR-testing in a representative random sample. PCR samples were collected in the period of April 1 to April 6, 2020. Between April 6 and April 10, the tested persons were contacted again by telephone to collect further information, e.g. on their state of health and mobility patterns.

**Methodology****Data collection period**

1st April 2020 - 10th April 2020

**Analysis unit**

Individual

**Country**

Austria

**Sampling procedure**

Probability: Stratified: Disproportional

**Time dimension**


Cross-section

**Data collection mode**

Physical measurements and tests; Telephone interview: CATI

Spread of COVID-19 in Austria. PCR-tests in a rep

Find [Advanced Search](#)

-  **Dataverses (14)**
-  **Datasets (150)**
-  **Files (112)**

1 to 10 of 276 Results

Sort ▾

**Dataverse Category**

- Organization or Institution (5)
- Journal (3)
- Research Project (3)
- Research Group (2)

**Publication Year**

- 2020 (126)
- 2019 (73)
- 2018 (69)
- 2017 (8)

**Author Name**

GfK (88)

**Spread of COVID-19 in Austria. PCR-tests in a representative sample (SUF edition)** 

May 14, 2020 - COVID-19 Pandemic



SORA Institute for Social Research and Consulting, 2020, "Spread of COVID-19 in Austria. PCR-tests in a representative sample (SUF edition)", <https://doi.org/10.11587/X2MIHW>, AUSSDA, V2, UNF:6:C2yOdfvGy+jqVa1CNoe88g== [fileUNF]

... is the first in continental Europe based upon nationwide **PCR-testing** in a **representative random sample**. **PCR** ...

Alternative Title: **COVID-19** Prävalenz in Österreich

Geographic Coverage Country / Nation: **Austria**

Keyword Term: **COVID-19**

Collection Mode: Physical measurements and **tests**; Telephone interview: CATI

Universe: All people living in **Austria** (excluding those currently in hospital), encompassing all age groups

**COVID-19 Pandemic (AUSSDA - The Austrian Social Science Data Archive)** 



Apr 30, 2020

Social Science Data on Coronavirus Disease Your information hub for SARS-CoV-2 / **COVID-19** pandemic

# Wyszukiwanie danych

Strony internetowe projektów badawczych, np.

- European Social Survey
- World Values Survey
- IPUMS: zharmonizowane międzynarodowe dane społeczno-demograficzne
- ParlGov: Parliaments and governments database: składy parlamentów i wyniki wyborów w krajach OECD
- GROWup: Geographical Research On War, Unified Platform: dane o grupach etnicznych/narodowościowych i konfliktach z nimi związanych
- Manifesto Project: pozycje partii politycznych w 50 krajach świata na podstawie manifestów partyjnych, od 1945 r.
- Varieties of Democracy (V-Dem): wskaźniki jakości demokracji
- ARDA (Association of Religious Data Archives): grupy religijne

# Wyszukiwanie danych

Dane zdeponowane w otwartych repozytoriach:

- Dataverse
- Open Science Framework
- Github

Dane zarchiwizowane poza internetem



# Cytowanie danych

Autor/producent/wydawca

Nazwa/tytuł

Data publikacji

Wersja/edycja/data wytworzenia

Forma zapisu

Identyfikator/link

Data Citation Principles:

<https://www.force11.org/datacitationprinciples>

DOI Citation Formatter: <https://citation.crosscite.org/>

Repozytoria danych sondażowych

# Europejski Sondaż Społeczny

European Social Survey

<http://www.europeansocialsurvey.org/>

<http://nesstar.ess.nsd.uib.no/webview/>

Formularz rejestracyjny:

<https://www.europeansocialsurvey.org/user/new>

# Światowe Badanie Wartości

World Values Survey:

<http://www.worldvaluessurvey.org/>

Zadanie domowe

# Analiza wybranych danych wtórnych

1. Pytanie badawcze
2. Opis danych (dostosowany do typu danych)
  - Rodzaj danych, źródło / sposób pozyskania
  - Adekwatność wybranych danych
  - Populacja, typ próby, poziom realizacji, rozmiar próby
  - Opis wybranych zmiennych; treść pytań i kategorie odpowiedzi
  - Problemy dot. jakości danych i ich wpływ na analizy
3. Analiza
  - Metoda analizy
  - Tabele / wykresy / schematy
4. Wnioski

Prezentacja 15-20 minut, 16 czerwca

Tekst 3-5 stron, do końca czerwca (mkolczynska@gmail.com)

# Linki

Europejski Portal Danych: <https://www.europeandataportal.eu/>

Otwieranie danych. Podręcznik dobrych praktyk:

<https://dane.gov.pl/media/ckeditor/2018/11/22/otwieranie-danych-podrecznik-dobrych-praktyk.pdf>

Listy repozytoriów danych społecznych:

<https://guides.lib.calpoly.edu/c.php?g=261997&p=1749797>

<https://guides.lib.vt.edu/c.php?g=580714>

Lista sondaży porównawczych: <https://www.gesis.org/en/services/data-analysis/more-data-to-analyze/overviews/overview-of-comparative-surveys-worldwide>

Cytowanie danych: <https://guides.lib.umich.edu/c.php?g=282964&p=3285995>

# Linki

Archiwa i repozytoria:

Archiwum Danych Społecznych: <http://www.ads.org.pl/>

Archiwum Danych Jakościowych: <http://www.adj.ifispan.pl/>

GESIS: <http://gesis.org>

UK Data Archive: <https://www.data-archive.ac.uk/>

ICPSR: <https://www.icpsr.umich.edu/>

Otwarte repozytoria:

Dataverse: <https://dataverse.harvard.edu/>

Open Science Framework: <https://osf.io/>

Zenodo: <https://zenodo.org/>

Figshare: <https://figshare.com/>



# Podziękowania



N A R O D O W E C E N T R U M N A U K I

Prezentowane treści powstały w ramach projektu „Uwarunkowania i konsekwencje zaufania politycznego: Polaryzacja polityczna i demokratyczna użyteczność zaufania w perspektywie porównawczej” (2019/32/C/HS6/00421).