BRILL

RESEARCH DATA JOURNAL FOR THE HUMANITIES AND SOCIAL SCIENCES 6 (2021) 1-16

BRILL OPEN

brill.com/rdj

The Longitudinal IntermediaPlus (2014–2016): A Case Study in Structuring Unstructured Big Data

Other Humanities

Inga Brentel (*corresponding author*) | ORCID: 0000-0002-9066-9191
Department for Communication and Media Studies, Institute of Social Science, Heinrich-Heine-University, Düsseldorf, Germany
inga.brentel@uni-duesseldorf.de

Kristi Winters
GESIS, Cologne, Germany
kristi.winters@gesis.org

Abstract

This article details the novel structure developed to handle, harmonize and document big data for reuse and long-term preservation. '*The Longitudinal IntermediaPlus (2014–2016)*' big data dataset is uniquely rich: it covers an array of German online media extendable to cross-media channels and user information. The metadata file for this dataset, and its documentation, were recently deposited as its own MySQL database called *charmstana_sample_14-16.sql* (https://data.gesis.org/sharing/#!Detail/10.7802/2030) (CS16) and is suitable for generating descriptive statistics. Analogous to the 'Data View' in SPSS, the *charmstana_analysis* (CA) contains the dataset's numerical values. Both the CS16 and CA MySQL files are needed to conduct analysis on the full database. The research challenge was to process large-scaled datasets into one longitudinal, big-data data source suitable for academic research, and according to FAIR principles. The authors review four methodological recommendations that can serve as a framework for solving big-data structuring challenges, using the harmonization software *CharmStats*.

Keywords

big data – big data documentation – data harmonization – German media data – longitudinal data – unstructured data

Online publication date: 6-7-2021

– Related data set "Meta-Information on the Sample of the Media-Analysis Data: The Longitudinal IntermediaPlus (2014–2016)" with DOI www.doi.org/10.7802/2030 in repository "GESIS"

1. Introduction

This article will explain the novel structure developed to handle, harmonize and document big data for reuse and long-term preservation. *The Longitudinal IntermediaPlus (2014–2016)* big data dataset is unique in its richness: it covers an array of German online media extendable to cross-media channels and information on the users.[1] These data are suitable for investigating, inter alia, media use (online and potentially offline), inequalities between social or geographic factors, routines of different (social) groups, media concentration tendencies, audience and market fragmentation in Germany or in comparison with Germany. The metadata file for this dataset and its documentation were recently deposited as its own MySQL database called *charmstana_sample_14-16.sql* (CS16) and are available for download from GESIS-Leibniz Institute for the Social Sciences (see Brentel et al., 2020). Similar to the 'Variable View' in SPSS, the CS16 database contains metadata on the full dataset and is suitable for generating an array of descriptive statistics. The CS16 file can be used to examine German media market analysis on the structural level, for example, the distribution of different German online media market genres (see Kampes, 2020) or the genre-portfolio of different media brands. Importantly, it also details our structuring solutions for the original, unstructured big data media files, including information extraction and the conceptual structure of the full database.

Analogous to the 'Data View' in SPSS, the *charmstana_analysis* (CA) is the MySQL database file containing the dataset's numerical values. Both the CS16 and CA MySQL files are needed to conduct analyses on The Longitudinal IntermediaPlus (2014–2016) database or extract variables of interest for analysis. The full CA database (<100GB) is embargoed until summer 2021, to be published with GESIS (current at the time of publication). However, prior to its full release in summer 2021, and upon e-mail request to the lead author, a chosen variable set of interest can be made accessible to researchers.

1 The mixed-methods design data collection includes some 100,000 cases for daily tracking of about 4,000 webpages, a combination of on-site and in-app questionnaires, and using a classic CATI-questionnaire survey is carried out twice a year. It excludes online media outlines under public law, such as ZDF and ARD.

Publications addressing the unique challenges for big data quality-standards have emerged in recent years (inter alia, Jürgens & Jungherr, 2016; van Atteveldt & Peng, 2018; van Atteveldt et al., 2019b; Wilkinson et al., 2016). This paper triangulates with the conceptual literature (Dienlin et al., 2021; Peter et al., 2020; RatSWD, 2019; RatSWD, 2020; van Atteveldt & Peng, 2018; van Atteveldt et al., 2019b; Wilkinson et al., 2016), the practical literature from the field of social media (Jürgens & Jungherr, 2016) and text analysis data (inter alia, Berman, 2013, pp. 2ff.; Lee et al., 2014) by documenting the lessons learned from our big data handling challenges and our technical solutions for big, semi-unstructured tracking and survey data. Four practical recommendations are provided in the conclusion that conform to scientific standards of transparency and reproducibility, following the FAIR principles of Wilkinson et al. (2016), and can be applied beyond text analysis and social media data.

2. Original Datasets: Description of the Big Data Dataset and the Research Problem

ag.ma's *IntermediaPlus* dataset combines digital trace data for online media use with representative survey data for the German population (over 14 years old). Due to rigorous operationalization by well-recognized, academic institutes for data collection (*cf.* Arbeitsgemeinschaft Media-Analyse e.V.,2020) high-quality data is produced. It includes, inter alia, information on cross-media use, on press media, radio, tv and online (Arbeitsgemeinschaft Media-Analyse e.V., 2014). The ag.ma IntermediaPlus data bundles have six Variable Sections (see Appendix). These bundles result from a joint venture of the German Media-Analysis agencies (ag.ma, agof and gfk/agf). It unites the media planning actors in the commercial branch, the broadcasting, and electronic media vendors branch. Each brought its perspective, different data needs and data use perspectives such as interpretation of media penetration.

Previously, these data were inaccessible because: 1) they are owned and embargoed for two years by companies (and although requests for the data for research were possible, access was not guaranteed); and 2) the structure is closed, it was sparsely documented and learning the data and the technical requirements demanded significant effort: a common issue when working with big data (Tekiner & Keane, 2013; van Atteveldt & Peng, 2018). The challenge of handling and tracking large amounts of metadata information became apparent quickly; because big data is broadly unstructured. The Variable Sections metadata in the original dataset were often identical: a single question but worded with different items, each with routine activity, a free-time activity or

media offering, with the same structure and metadata information, year on year (e.g., identical question wording and variable values multiplied the total number of variables significantly). With 4,000-plus online media-use variables to process for each year, all with identical metadata, we needed an automatized looping code that also produced aggregated data-documentation sufficiently detailed to conform to the FAIR principles (Mons et al., 2017, pp. 51f.). Similarly, in the Variable Section changes in the respondent's belongings, socio-demographic and household needed to be tracked and made visible for users.

3. Method: Planning and Digitizing the Workflow

Big data transformation requires advanced planning. Before this project, these data were stored as a large-scale, semi-unstructured data source, also known as a *data silo*, which is closed in its own storage structure and inaccessible to researchers. We pooled and transformed the Media-Analysis commercial media-market data source into a structured big-data dataset, with complete documentation, into a harmonized dataset called The Longitudinal IntermediaPlus Data Source (2014–2016). A customized 'per variable' tracking system was designed and documented the harmonization process via CharmStats. The Pressmedia and Radio bundle of Media-Analysis were harmonized in the same traceable and sustainable way (Brentel & Jandura, 2018, 2021; Jandura & Brentel, 2021; Jandura et al., 2021).

Our data processing solution was the use of automated "loops" developed by the lead author for use in the open-source variable harmonization software CharmStats. It has the capacity to generate some re-coding languages and fetch all the metadata associated with the project into a digital report. This allowed us to generate a user-friendly output table for a variable, with all relevant metadata information, across time, and displaying any changes (see Figure 5). These detailed data documents, with information informed by the FAIR principles, resulted in a highly-reusable, high-quality database.

We developed four steps to solve big-data processing challenges when working with categorical and metric variables for different years, that are large-scaled and relatively unstructured data (see Figure 1):

- Step 1: Plan out a codebook documentation structure strategy, conforming to the FAIR principles of Findability and Accessibility. This documentation structure strategy guided the shape of the files.
- Step 2: Identify the metadata for the study, question and variable levels to be imported into CharmStats from *.sav file or Open Office spreadsheets

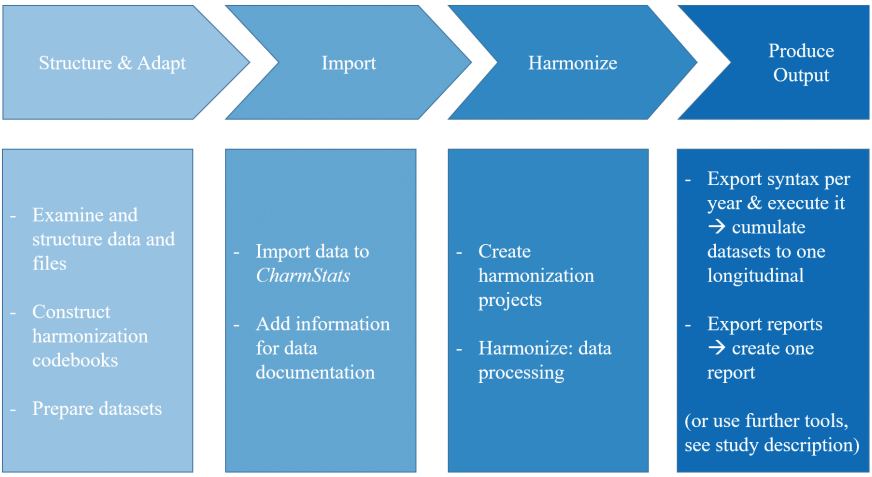

FIGURE 1 Four steps of success to produce the IntermediaPlus 2014–2016 longitudinal dataset

importation, which can be enriched by hand-entered information (e.g., bibliographical information and notes documenting harmonization decisions).
– Step 3: Complete the variable harmonization and data processing work in CharmStats.
– Step 4: Use CharmStats to produce outputs including the dataset documentation report, the recoding language for data processing in statistical software and codebook reports with complete data documentation, all conforming to the FAIR principles of Interoperability and Reproducibility.

4. Creating a Big Data Structure

Social scientists may view big-data sources as quite attractive. They are often free and rich sources of data. However, would-be big-data users face unique data handling challenges before they can do any data analysis (Japec et al., 2015; Jungherr et al., 2018, pp. 255f.). These data silos are usually stored as unstructured, large-scale big-data collections, requiring substantial data handling before they can be used (Foster et al., 2017, pp. 7ff.; Maroto, 2016; van Atteveldt & Peng, 2018). For those researchers who want to work with unstructured or semi-unstructured big data formats, but who are used to structured datasets, we propose the use of these data handling and documentation standards.[2]

2 For information on a big-data data quality framework from data science, see Cai & Zhu, 2015. The FAIR principles match this framework.

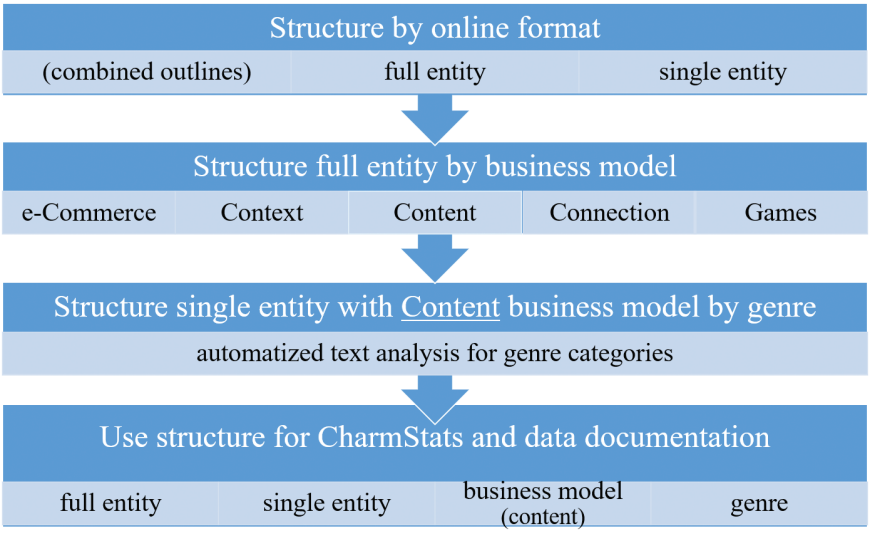

FIGURE 2 The structuring levels for the IntermediaPlus dataset

Our conceptual approach included first identifying important information in the unstructured or semi-unstructured data for *information extraction* (*cf.* Warin & Sanger, 2014).[3] We used this identified information to re-organize the unstructured big data into an understandable and re-usable structured big data dataset. Big data can be organized in a variety of ways, but our structuring was driven by which structures would best *answer our research questions*, while also adding secondary reuse value for the research community.[4] We structured these big data on different data "levels", as set out in Figure 2.

– The first structuring level reflects the practical reality of the existing data structure itself, namely *Full Entity*, *Single Entities*, and *Combined Outlines*.
– The second-order structuring level was conceptual: the typology for business models following Wirtz (2018, pp. 307ff.).
– Our third-order, and final, structuring level was genres (see Kampes, 2020).

These structures reduced the number of variables down to about half the number of relevant entities. The structures also guided the separating of different online entities into smaller groups, and those smaller groups facilitated processing metadata in CharmStats, while simultaneously enriching the data with more information by automatically adding para-data for later filtering or

3 Information extraction is a methodological approach used in computer science; see, for example, Gudivada et al., 2017, p. 5.

4 For more information, see the study description and documentation archived at GESIS.

analysis. This can be achieved with the structure as indicated by the variable names (see Brentel et al., 2020: data-documentation, Description of the work carried out). By way of example, the variables for full entity online offerings begin with the prefix "GA_" (Gesamt-Angebot) while those of a single entity start with "EA_" (Einzel-Angebot), and genre-category labels are displayed with a hashtag, for example, "#Digital".

5. The CharmStats Workflow

CharmStats is a software solution that offers a structured workflow to overcome big data challenges.[5] Developed at GESIS Leibniz Institute for the Social Sciences, it breaks down data processing into a metadata-based workflow. Built from DDI metadata standards, CharmStats stores metadata on:

- levels of the study, such as study name, collection dates, collection area;
- question, such as multi-lingual question wording, show card response options;
- variable including response options and corresponding labels; and
- value, adding a comment to a response value.

Users import metadata using SPSS or Open Office spreadsheets, then connect Source and Target Variables metadata in the MySQL database via interactive interfaces. The metadata organization in CharmStats, as shown in Figure 3, allows information to be reused in efficient ways. Users can connect, save, and retrieve metadata connections to generate individual digital documents or to produce codebooks making them reproducible.[6] The workflow traces a user's work to generate individualized data documentation outputs, including generating graphs that visualize the recoding structure and auto-generates code in SPSS, Stata, MPlus and SAS. To accommodate our project's data needs, CharmStats was adapted to include large-scale data processing for categorical and metric variables. Figure 4 presents an overview of the CharmStats workflow.

To manage the approximately 21,500 variables in the original ag.ma's IntermediaPlus datasets, first a Variable Sections list (e.g., Media Use, Free Time Activities, Socio-demographics) and then a list of variables per Section

5 The Coding and Harmonization of Statistics software CharmStats is a Java-based, open-source, free software that stores to, retrieves from and connects across harmonization metadata via a MySQL database.

6 In addition to big datasets used, CharmStats was used to document code for a crowdsourced replication initiative experiment (Breznau et al., 2019).

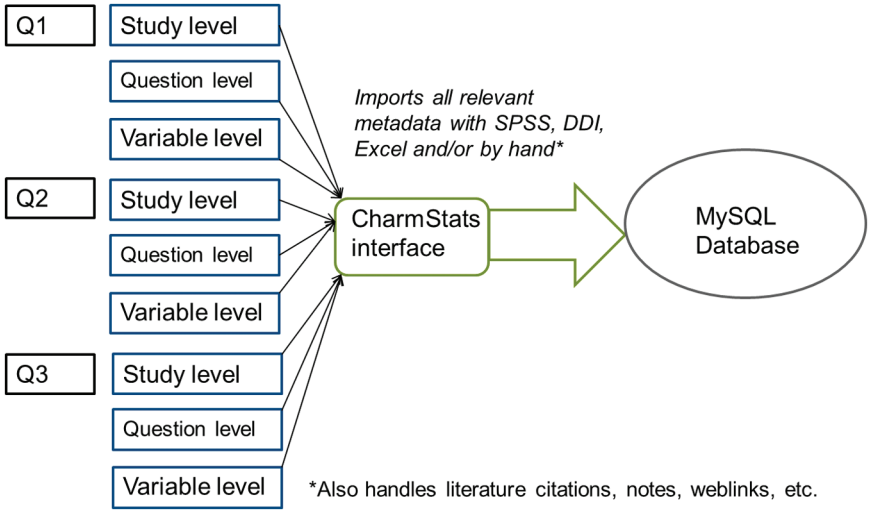

FIGURE 3 Representation of types of metadata handled by CharmStats

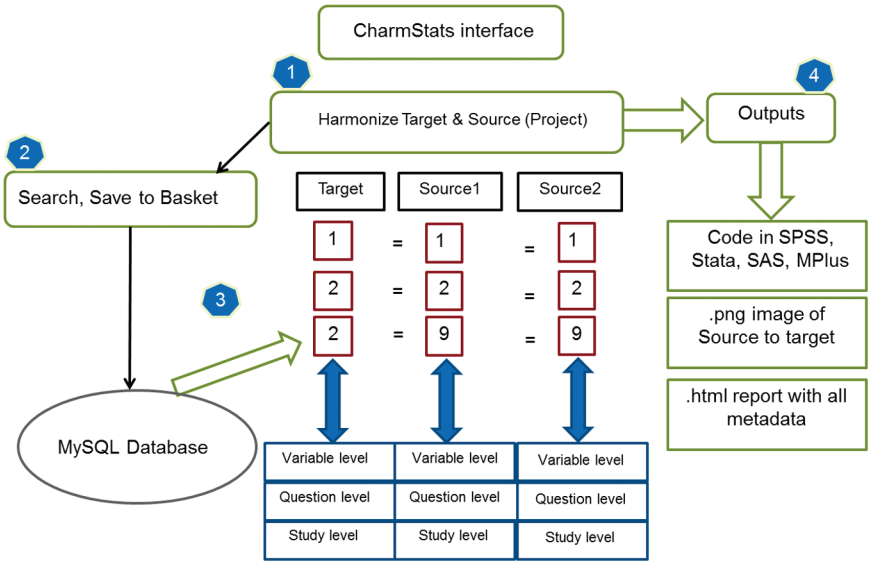

FIGURE 4 Conceptual representation of the CharmStats workflow and outputs

(e.g., hh_aboalice, education, GA_alst_CT_pi) were needed to structure the dataset. Using CharmStats, we generated digital documentation for our harmonization process using bespoke report templates based on our reporting needs. These digital reports tracked variable changes across the years as detailed documentation, per harmonized variable (as shown in Figure 5). The

ONLINE-MEDIENNUTZUNG CONTENT

Variablenname: GA_alst_CT_pi - Label: Page Impressions Alster Radio, Gesamtangebot: Content (if)
Variablenname: GA_lue_CT_pi - Label: Page Impressions Leben und Erziehen, Gesamtangebot: Content (if)
Variablenname: GA_alst_CT_nt - Label: Nutzung tagesbasis Alster Radio, Gesamtangebot: Content (if)
Variablenname: GA_lue_CT_nt - Label: Nutzung tagesbasis Leben und Erziehen, Gesamtangebot: Content (if)
Variablenname: EA_splt_DP_pi - Label: Page Impressions spieletipps - PC #Digital (if)
Variablenname: EA_splt_DP_nt - Label: Nutzung tagesbasis spieletipps - PC #Digital (if)
Variablenname: EA_bild_P_pi - Label: Page Impressions BILD - Politik #Politik (if)
Variablenname: EA_bild_P_nt - Label: Nutzung tagesbasis BILD - Politik #Politik (if)

Variablenname: GA_alst_CT_pi - Label: Page Impressions Alster Radio, Gesamtangebot: Content (if)

Variablenname: GA_7fm_CT_pi - Label: Page Impressions 7forum, Gesamtangebot: Content (if)
Variablenname: GA_7tv_CT_pi - Label: Page Impressions 7TV / joyn, Gesamtangebot: Content (if)
Variablenname: GA_90m_CT_pi - Label: Page Impressions 90Min, Gesamtangebot: Content (if)
[…]

Frage: F7D: Wenn II2 nur Computer=1-6 Ich nenne Ihnen jetzt eine Auswahl verschiedener Webseiten, die man im Internet finden kann. Wenn II2 Tablet=1-6 und/oder Smartphone=1-6 Ich nenne Ihnen jetzt eine Auswahl verschiedener Webseiten oder #Apps#, die man im Internet finden kann. Welche dieser Webseiten oder #Apps# haben Sie in den letzten 3 Monaten besucht oder genutzt, egal ob [Orte aus F2 werden eingespielt, z.B.: „zu Hause, am Arbeitsplatz bzw. am beruflichen Ausbildungsplatz"] oder an einem anderen Ort. Ab dem 2. Item: Haben Sie [Angebot] in den letzten 3 Monaten besucht oder genutzt? F7E: Wenn F7D für JEDES Angebot = 2 oder NULL oder wenn F7D lediglich Google und/oder Facebook = 1 Abgesehen von den Angeboten, von denen wir gerade gesprochen haben, nennen Sie mir doch bitte bis zu drei weitere Internet-Angebote, die Sie in den letzten 3 Monaten genutzt haben

Kodierung:
-7, nicht ermittelt

Intervieweranweisung: F7D: INT: Bei Rückfragen: Auch #Apps# sind gemeint (Liste der Angebote randomisiert) AOL DE boerse #online# DE brigitte DE #chip online# #cinema# DE #ebay# DE eins und eins DE eltern DE falk DE #facebook# finanztreff DE fit #for fun# DE focus #online# #freenet# DE freundin DE #gamestar# DE GMX #google# handelsblatt com meinestadt DE mobile DE MSN DE NTV DE PC welt DE prosieben DE RTL DE spiegel #online# sport eins #stayfriends# DE suddeutsche DE T #online# DE TV spielfilm #web# DE wetter com wirtschaftswoche DE #yahoo# die zeit DE zeit #online# Ja 1 --> F7Da Nein 2 --> nächstes Angebot mögliche Antwort NULL oder WN F7E: mögliche Antwort NULL oder WN

Vermerk: Die Harmonisierung dieser Variable steht stellvertretend für die Harmonisierung aller Variablen dieser Gruppe. Diese haben die gleichen Variableninformationen bezüglich der Kodierung, Variablenbildung, etc. und wurden daher als virtuelle Variablen aufgenommen.

Übersichtstabelle der Variable und Harmonisierung für die Jahre 2014 bis 2016

	2014	2015	2016
Variablenname	G10021, G1011, G1026, G1031, G1036, G1051, G1061, G1066, ***[…]***	G1001, G1006, G1011, G1016, G1026, G1036, G1046, G1051, ***[…]***	G11, G1326, G1331, G1336, G1346, G13506, G1351, G1356, ***[…]***
Frage	F7D: Wenn II2 nur Computer=1-6 Ich nenne Ihnen jetzt eine Auswahl verschiedener Webseiten, die man im Internet finden kann. ***[…]***	F7D: FILTER: WENN II2 nur Computer=1-6 Ich nenne Ihnen jetzt eine Auswahl verschiedener Webseiten, die man im Internet finden kann. ***[…]***	F7D: FILTER: WENN II2 Computer = 1-6 UND/ODER Tablet=1-6 UND/ODER Smartphone=1-6 Ich nenne Ihnen jetzt eine Auswahl verschiedener Webseiten oder #Apps#, die man im Internet finden ***[…]***
Intervieweranweisung	F7D: INT: Bei Rückfragen: Auch #Apps# sind gemeint (Liste der Angebote randomisiert) AOL DE boerse #online# DE brigitte DE #chip online# #cinema# DE #ebay# DE ***[…]***	F7D: INT: Bei Rückfragen: Auch #Apps# sind gemeint (Liste der Angebote randomisiert) AOL boerse #online# DE brigitte DE #chip online# #cinema# #ebay# eins und ***[…]***	F7D: INT: Bei Rückfragen: Auch #Apps# sind gemeint (Liste der Angebote randomisiert) AOL boerse #online# DE brigitte DE #chip online# #cinema# #ebay# eins und ***[…]***
Kodierung	-7, nicht ermittelt	-7, nicht ermittelt	-7, nicht ermittelt
Harmonisierungs-Syntax	RECODE (-7 = -7) (ELSE = COPY) INTO GA_alst_CT_pi. MISSING VALUES GA_alst_CT_pi (-7). EXECUTE.	RECODE (-7 = -7) (ELSE = COPY) INTO GA_alst_CT_pi. MISSING VALUES GA_alst_CT_pi (-7). EXECUTE.	RECODE (-7 = -7) (ELSE = COPY) INTO GA_alst_CT_pi. MISSING VALUES GA_alst_CT_pi (-7). EXECUTE.
Vermerk	Die Frageformulierung basiert jeweils auf dem Fragebogen der 2. Erhebungswelle.	Die Frageformulierung basiert jeweils auf dem Fragebogen der 2. Erhebungswelle.	Die Frageformulierung basiert jeweils auf dem Fragebogen der 2. Erhebungswelle.

Note: Based on data documentation in Brentel et al. (2020).

FIGURE 5 Sample of individualized data documentation

reports included researcher comments, replication information for later users, hyperlinks within the document so users need not scroll through 18,000 variables of the final data-documentation, thereby making it user-friendly and transparent and facilitating future replication.

6. Resulting Data: the Longitudinal IntermediaPlus Data Source (2014–2016)

- **The Longitudinal IntermediaPlus deposited at GESIS** – DOI:www.doi.org/10.7802/2030
- **Temporal coverage:** 2014–2016

These cross-sectional datasets were pooled and transformed into one longitudinal, big-data source, The Longitudinal IntermediaPlus (2014–2016) (<100GB). The metadata file for this dataset, and its documentation, are available for download from GESIS as its own MySQL database called *charmstana_sample_14-16.sql* (CS16). The deposited CS16 metadata and its data documentation detail the exact shape of the variables. It facilitates metadata analysis on the German media market and enables users of the CA to compile their variable 'set of interest' to create their own, bespoke version of the full MySQL database for analysis.

The *charmstana_analysis* (CA) is the MySQL database file containing the dataset's numerical values. The forthcoming CA MySQL database has a specialized structure to facilitate complex computational analysis and efficient computer performance despite its big size. The pooled and longitudinal CA dataset has around 18,000 variables with more than 1.6 million German respondents. These data are suitable for investigating, inter alia, media use (online and potentially offline), inequalities between social or geographic factors, routines of different (social) groups, media concentration tendencies, audience and market fragmentation in Germany or in comparison with Germany. Variables included are socio-demographic characteristics, free-time activities, the respondent's belongings and social class. There are variables for online media market characteristics, for example, media provider, marketer, genre, business model or origin of an online media offering. Geographical variables allow analysis on the level of governing districts. The high number of cases in the dataset, and it being statistically representative for Germany, enables complex statistical methods, such as network analysis and analysis for specific (sub) groups without problems of a small-N (van Atteveldt & Peng, 2018, pp. 83f.;

van Atteveldt et al., 2019a, p. 3). Both the CS16 and CA MySQL files are needed to conduct analysis on the full database, or extract variables of interest for analysis. The published metadata and data documentation in the CS16 are needed to understand the CA MySQL database and prepare the query for the researcher's own, a customized version of this database.

7. Conclusion

We confronted the challenge of harmonizing large-scale, semi-unstructured and cross-sectional data silos covering several years, a data preparation challenge that future researchers will also have to confront. We found that digitizing the repetitive work of recoding and documenting using CharmStats was very useful in this large-scale data project. The transformed dataset is now pooled, clearly-structured, longitudinal, readable and understandable, and it comes with finable, accessible, interoperable and reusable documentation ready for academic use. The inclusion of precise recoding instructions for SPSS, STATA, MPlus and SAS are of particular value to researchers.

To replicate this process, or apply it to another data silo, we offer four recommendations:

(1) use the inherent data structures and decide a conceptual structure to match your research interest;
(2) import the relevant metadata into CharmStats using SPSS or an Open Office spreadsheet program;
(3) harmonize the metadata as per your pre-defined structure, make use of the CharmStats features, such as automatized data processing; and
(4) produce your outputs and use a template option to export your bespoke documentation.

For more information on the Longitudinal IntermediaPlus (2014–2016) or the longitudinal datasets for Radio (1977–2015) and Pressmedia (1954–2015) visit the GESIS website for Media-Analysis data or email the lead author, Inga Brentel.

Acknowledgements

This work was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia. The original data as well as information on data collection was kindly provided by ag.ma and agof.

References

Arbeitsgemeinschaft Media-Analyse e.V. (2014). *Datensatz Codeplan MA 14*.

Arbeitsgemeinschaft Media-Analyse e.V. (2020). Datenerhebung der ma Intermedia PLuS. https://www.agma-mmc.de/media-analyse/ma-intermedia-plus/datenerhebung.

Berman, J. J. (2013). *Principles of big data: Preparing, sharing, and analyzing complex information*. *Safari Tech Books Online*. Morgan Kaufmann.

Brentel, I., & Jandura, O. (2018). *Media-Analyse: Radio – Langfristdaten*. https://www.doi.org/10.7802/1620.

Brentel, I., & Jandura, O. (2021). *Media-Analyse: Pressemedien – Langfristdaten* (*Version 2.0*). https://www.doi.org/10.7802/2157.

Brentel, I., Kampes, C. F., & Jandura, O. (2020). *Meta-Information des Samples der Media-Analyse Daten: IntermediaPlus* (*2014–2016; Version: 1.0.0*). SoWiDataNet. GESIS. https://www.doi.org/10.7802/2030.

Breznau, N., Rinke, E. M. & Wuttke, A. (2019). *OSSC19 Crowdsourced Replication Initiative*, Mannheim Centre For European Social Research (MZES), University of Mannheim. https://harmonization.gesis.org/#!BrowseResults/?searchval=Crowdsource.

Cai, L. & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the Big Data Era. *Data Science Journal*, 14, p.2. http://www.doi.org/10.5334/dsj-2015-002.

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., Lukito, J., Bier, L. M., Zhang, R., Johnson, B. K., Huskey, R., Schneider, F. M., Breuer, J., Parry, D.A., Vermeulen, I., Fisher, J.T., Banks, J., Weber, R., Ellis, D.A., … de Vreese, C. (2021). An agenda for open science in communication. *Journal of Communication*, 71(1), 1–26. https://www.doi.org/10.1093/joc/jqz052.

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F. & Lane, J. (2017). *Big data and social science: A practical guide to methods and tools*. CRC Press.

Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, 10, 1–20.

Jandura, O., & Brentel, I. (2021, forthcoming). *Media-Analyse-Daten: Radio-Tranche* (*2010–2015; MA-Radio*). GESIS Datenarchiv, Köln. ZA5762 Datenfile Version 1.0.0, doi:https://www.doi.org/10.4232/1.13662.

Jandura, O., Brentel, I., & Babic, D. (2021). *Media-Analyse-Daten: Pressemedien-Tranche* (*2010–2015*). GESIS Datenarchiv, Köln. ZA5761 Datenfile Version 1.0.0. https://www.doi.org/10.4232/1.13661.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., & Usher, A. (2015). Big Data in survey research. *Public Opinion Quarterly*, 79(4), 839–880. https://www.doi.org/10.1093/poq/nfv039.

Jungherr, A., Jürgens, P., & Schoen, H. (2018). 12 Twitter-Daten in der Wahlkampfforschung: Datensammlung, Aufarbeitung und Analysebeispiele. In A. Blätte, J. Behnke, K.-U. Schnapp, & C. Wagemann (Eds.), *Schriftenreihe der Sektion Methoden der Politikwissenschaft der Deutschen Vereinigung für Politische Wissenschaft. Computational Social Science: Die Analyse von Big Data* (1st Ed., pp. 255–294). Nomos Verlagsgesellschaft.

Jürgens, P., & Jungherr, A. (2016). A tutorial for using Twitter data in the social sciences: Data collection, preparation, and analysis. *SSRN Electronic Journal.* Advance online publication. https://www.doi.org/10.2139/ssrn.2710146.

Kampes, C. F. (2020). Welche Genres existieren für Online-Medienangebote? Eine Analyse der Themenstruktur aus Anbietersicht. In W. Deiters, S. Geisler, F. Hörner, & A. K. Knaup (Eds.), *Die Kommunikation und ihre Technologien. Interdisziplinäre Perspektiven auf Digitialisierung* (pp. 13–44). Transcript Verlag.

Lee, K., Noh, Y., Yoon, S., & Cho, Y. (2014). Structuring of unstructured big data and visual interpretation. *Journal of the Korean Data and Information Science Society*, 25(6), 1431–1438. https://www.doi.org/10.7465/jkdi.2014.25.6.1431.

Maroto, C. (2016). *A data lake architecture with Hadoop and open source search engines: Using enterprise data lakes for modern analytics and business intelligence*. Retrieved from https://www.dzone.com/articles/a-data-lake-architecture-with-hadoop-and-open-sour.

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, 37(1), 49–56. https://www.doi.org/10.3233/ISU-170824.

Peter, C., Breuer, J., Masur, P. K., Scharkow, M., & Schwarzenegger, C. (2020, December 11). *Empfehlungen zum Umgang mit Forschungsdaten in der Kommunikationswissenschaft*. Retrieved from https://www.dgpuk.de/sites/default/files/AG_Forschungsdaten%20Empfehlungen%20DGPuK_0.pdf.

Rat für Sozial- und Wirtschaftsdaten [RatSWD] (2019). *Big Data in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Datenzugang und Forschungsdatenmanagement – Mit Gutachten "Web Scraping in der unabhängigen wissenschaftlichen Forschung"* (Output No. 4[6]). https://www.doi.org/10.17620/02671.39.

Rat für Sozial- und Wirtschaftsdaten [RatSWD] (2020). *Datenerhebung mit neuer Informationstechnologie: Empfehlungen zu Datenqualität und -management, Forschungsethik und Datenschutz* (Output No. 6[6]). https://www.doi.org/10.17620/02671.47.

Tekiner, F., & Keane, J. A. (2013, October 13–16). Big Data framework. In *Proceedings: 2013 IEEE International Conference on Systems, Man and Cybernetics: SMC 2013, Manchester, United* Kingdom (pp. 1494–1499). IEEE Computer Society. https://www.doi.org/10.1109/SMC.2013.258.

van Atteveldt, W., Margolin, D., Shen, C., Trilling, D., & Weber, R. (2019a). A roadmap for computational communication research. *Computational Communication Research*, 1(1), 1–11. https://www.doi.org/10.5117/CCR2019.1.001.VANA.

van Atteveldt, W., & Peng, T.Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls, *Computational Communication Science. Communication Methods and Measures*, 12(2–3), 81–92. https://www.doi.org/10.1080/19312458.2018.1458084.

van Atteveldt, W., Strycharz, J., Trilling, D., & Welbers, K. (2019b). Toward open computational communication science: A practical road map for reusable data and code. *International Journal of Communication*, 13, 3935–3954.

Warin, T., & Sanger, W. (2014). Structuring big data: How financial models may help. *Journal of Computer Science and Information Technology*, 2, 1–20.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, Article 160018. https://www.doi.org/10.1038/sdata.2016.18.

Wirtz, B. W. (2018). *Electronic business* (6th Ed.). Springer.

Appendix

TABLE A1 Variable Sections of Intermedia Plus

Variable Section	Variable example	Challenge	Number of variables
Socio-demographic	Age, education, employment and income as well as the respondents' household (e.g. number of children, respondents' origin, and local population)	Changes over time, case sensitive, data documentation	20
Free time activities		Same structured; looping as well as aggregated data-documentation is helpful and increases clarity	10
Respondent belongings		(Varies in number), changes over time, esp. content wise, data documentation helps to see social changes through items asked immediately	18
Belongings of the household		(Varies in number), changes over time, esp. content wise, data documentation helps to see social changes through items asked immediately	11

TABLE A1 Variable Sections of Intermedia Plus (*cont.*)

Variable Section	Variable example	Challenge	Number of variables
Media use	About 100 radio stations; more than 150 magazines and roughly 100 newspapers; use of approximately 8,000 webpages (as measured by page impressions, as measured by average week or month, and also quarterly which was the original metric used by AGOFS on a daily basis); at least 10 tv-channels	Same structured; looping as well as aggregated data-documentation is essential and increases clarity (It is the same structure within a media use bundle or variable set → challenge: large-scaled as same structure gets multiplied, esp. for radio measured hourly)	Roughly 21,000 available variables for media usage (about 4,000 radio, 295, pm, 16,000 online, 445 tv)
Other	variables regarding the interviews like the date, the respondents' interest in the interview, etc. and weighting variables	Relatively stable, no need for harmonization as highly case sensitive	~10