Katalog archivischer Dateiformate KaD
Statistikformate: Situationsanalyse und Perspektiven

1 Anwendungen und Formate

Unter Statistikformaten verstehen wir im Folgenden Dateiformate, die für die Speicherung von statistischen Daten verwendet werden. In diesem Bereich dominieren die proprietären Formate der wichtigen Statistiksoftwarelösungen SPSS, Stata und SAS.

SPSS, mit vollem Namen IBM SPSS Statistics, wurde ursprünglich für die Sozialwissenschaften entwickelt und wird immer noch stark im wissenschaftlichen Bereich verwendet. Neben dem nativen Dateiformat .sav kann SPSS Daten im proprietären *SPSS Portable file format* .por speichern, welches als Austauschformat sehr weit verbreitet ist.

Stata ist eine Software für Datenanalyse und Statistik. Ihr proprietäres Dateiformat *Stata_dta* mit der Dateiendung .dta ist auch als Austauschformat sehr weit verbreitet.

SAS ist eine Softwaresuite für die statistische Analyse und stammt ursprünglich aus dem akademischen Umfeld. Das *SAS Transport File Format* (oder SAS_xport) ist ein proprietäres, aber öffentlich dokumentiertes Austauschformat, das vor allem in der Pharmaindustrie sehr verbreitet ist.

Als Austauschformat für statistische Daten ist auch *CSV* relativ verbreitet. Seltener ist der Gebrauch von *Microsoft Excel-Dateien*. Als mögliches zukünftiges Archivformat steht das offene Austauschformat *SDMX* (Statistical Data and Metadata eXchange, https://sdmx.org/, ISO 17369:2013) im Raum, welches von einer Community internationaler Institutionen hauptsächlich aus der Finanz- und Wirtschaftswelt entwickelt wird. Dieses Format scheint allerdings in den Datenarchiven noch kaum angekommen zu sein.

2 Best Practice

Ein Survey bei 8 wichtigen Datenarchiven in den Sozialwissenschaften zeigt einen weitgehenden Konsens für eine *Best Practice* zur Archivierung von Statistikdaten. Die Austauschformate der drei wichtigen Statistikprogramme SPSS (.por), STATA (.dta) und SAS (.sas) sind in dieser Reihenfolge praktisch überall akzeptiert, .por sogar überall. Teilweise sind auch die proprietären Formate von SPSS (.sav) und SAS (.sas7bdat) akzeptiert. Das Statistikprogramm R bzw. sein Format spielt nur eine untergeordnete Rolle. Bei den offenen Formaten sind die verschiedenen Ausprägungen von CSV breit akzeptiert, in der Regel mit zusätzlichen Metainformationen. Excel-Dateien oder Datenbankformate wie MDB werden auch hin und wieder akzeptiert. Zu beachten ist, dass diese Institutionen explizit auf die Nachnutzung der Daten ausgelegt sind.

Die Policies der folgenden Institutionen wurden analysiert:

- *FORS (CH), nationales Kompetenzzentrum für die Sozialwissenschaften, http://forscenter.ch/de/daris-daten-und-forschungsinformationsservice/datenservice/daten-hinterlegen/richtlinien-fur-das-hinterlegen-von-daten/vorbereiten-von-quantitativen-daten/*
- *GESIS (D), Leibniz-Institut für Sozialwissenschaften, http://www.gesis.org/angebot/archivieren-und-registrieren/datenarchivierung/vorbereitung-datenuebergabe/*

- *ICPSR (USA), Institute for Social Research, University of Michigan, https://www.icpsr.umich.edu/icpsrweb/content/shared/ICPSR/faqs/what-kind-of-data-formats-does-archive.html*
- *SSCS (USA), Social Sciences Computing Services, University of Chicago, https://sscs.uchicago.edu/page/data-archive-documentation*
- *Dataverse Project (USA), The Institute of Quantitative Social Science, Harvard, http://guides.dataverse.org/en/4.6.1/user/*
- *Social Science Data Archive at UCLA (USA), http://data-archive.library.ucla.edu/SSDA_collectionAndArchivingPolicy.pdf?_ga=1.251788362.2115979802.1490092210*
- *UK Data Service, University of Essex, https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats*
- *Australian Data Archive, https://www.ada.edu.au/ada/preferred-formats*

Archiv	*SPSS .por*	*SPSS .sav*	*STATA .dta*	*SAS .sas*	*SAS .sas7bdat*	*R*	*CSV*	*XLSX*
FORS	X	(X)						
GESIS	X	X	X	X	X		X	(X)
ICPSR	X	X	X	X			X	
SSCS	X	X	X	X	X			
Dataverse	X	X	X			X	(X)	X
UCLA	X		X	X			(X)	
UKDS	X	(X)	(X)		X		X	(X)
ADA	X	X	X	(X)		(X)	X	(X)

3 Empfehlung

Statistikdaten spielen in den KOST-Trägerarchiven noch kaum eine Rolle. Eine vertiefte Untersuchung der möglichen Dateiformate ist deshalb momentan nicht prioritär. Falls den Archiven Statistikdaten angeboten werden, ist es empfehlenswert, sich an der *Best Practice* von spezialisierten Institutionen zu orientieren und als Dateiformat eines der weit verbreiteten Austauschformate zu fordern. Am weitesten verbreitet ist .por, vorzuziehen ist allenfalls das offengelegte *SAS transport file format*. Die Verbreitung des ISO-Standards *SDMX* scheint noch zu gering, um dessen Verwendung im Archivbereich empfehlen zu können.

Stand: 25.07.2017