Catalogue des formats de fichiers d'archivage (Cfa)
Formats statistiques :
Analyse de la situation et perspectives

1 Applications et formats

Nous entendons par formats statistiques les formats de fichiers suivants utilisés pour enregistrer des données statistiques. Dans ce domaine, les formats propriétaires des principales solutions logicielles statistiques SPSS, Stata et SAS ont une position dominante.

SPSS, dont le nom complet est IBM SPSS statistics, a été développé à l'origine pour les sciences sociales et il est encore très utilisé dans le domaine scientifique. En plus du format de fichier d'origine .sav, SPSS peut stocker des données dans le format d'échange très répandu SPSS *Portable file format* .por.

Stata est un logiciel pour l'analyse de données et la statistique. Son format de fichier propriétaire *Stata_dta* avec l'extension .dta est également très répandu comme format d'échange.

SAS est une suite logicielle pour l'analyse statistique et provient à l'origine du monde académique. Le *SAS Transport File Format* (ou SAS_xport) est un format d'échange propriétaire, mais qui bénéficie d'une documentation publique. Il est avant tout très répandu dans l'industrie pharmaceutique.

CSV est également relativement répandu comme format d'échange pour données statistiques. Des fichiers *Microsoft Excel* sont utilisés plus rarement. Il est question du format d'échange ouvert *SDMX* (Statistical Data and Metadata eXchange, https://sdmx.org/, ISO 17369:2013) comme format d'archivage potentiel à l'avenir. Ce format est développé par une communauté d'institutions internationales issues principalement du monde de l'économie et de la finance. Il semble qu'il soit toutefois très peu parvenu dans les archives de données.

2 Bonnes pratiques

Un aperçu de 8 importantes archives de données dans le domaine des sciences sociales montre un large consensus sur des bonnes pratiques en matière d'archivage de données statistiques. Les formats d'échange des trois grands programmes statistiques SPSS (.por), STATA (.dta) et SAS (.sas) sont acceptés dans cet ordre pratiquement partout, et même partout en ce qui concerne .por. Les formats propriétaires de SPSS (.sav) et SAS (.sas7bdat) sont parfois également acceptés. Le programme statistique R, respectivement son format, ne joue qu'un rôle secondaire. Parmi les formats ouverts, les différentes formes de CSV sont largement acceptées, en règle générale avec des méta-informations supplémentaires. Des fichiers Excel ou des formats de bases de données comme MDB sont également acceptés parfois. Notons que ces institutions sont explicitement conçues pour l'utilisation de ces données.

Les politiques des institutions suivantes ont été analysées :

- *FORS (CH), centre de compétences suisse en sciences sociales, https://forscenter.ch/wp-content/uploads/2018/10/collections-policy.pdf*

- *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*

Archives	*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 Recommandation

Les données statistiques ne jouent qu'un rôle secondaire dans les archives membres du CECO. Un examen approfondi des formats de fichiers potentiels n'est donc pas prioritaire pour l'instant. Si des archives statistiques sont proposées aux services d'archives, nous recommandons de suivre les bonnes pratiques d'institutions spécialisées et d'exiger comme format de fichier l'un des fichiers d'échange largement répandus. Le plus répandu de tous est .por, le cas échéant, il convient de privilégier le format ouvert *SAS transport file format.* La norme ISO *SDMX* ne semble pas encore assez répandue pour que nous puissions recommander son utilisation dans le domaine archivistique.

État : 25.07.2017, liens mis à jour en octobre 2021