Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Bundesamt für Statistik BFS
Abteilung Kommunikation und Publishing

Diffusion von statistischen Ergebnissen

Richtlinien und Regeln für die Produktion von Dateien im CSV-Format beim BFS

14.07.2023 - Version 03

Wichtig: Das vorliegende Dokument bezieht sich ausschliesslich auf Dateien mit statistischen Ergebnissen, d.h. mit statistischen Zahlenangaben, die sich aus der Kombination mehrerer Variablen ergeben.

Die Richtlinien gelten **nicht** für CSV mit Metadaten oder Nomenklaturen.

Bei Dateien mit Klassifikationen oder Kodierungen wie beispielsweise der Schweizerischen Operationsklassifikation (CHOP) müssen die Regeln und Empfehlungen in diesem Dokument folglich nicht berücksichtigt werden.

CSV-Dateien, die aus den BFS-Anwendungen für die Datendiffusion exportiert werden, weichen unter Umständen von den hier genannten Richtlinien ab (STAT-TAB usw.).

Inhaltsverzeichnis

1. Dateiformate

Im Rahmen der Open-Government-Data-Strategie (OGD) des Bundesrates möchte das BFS die Bandbreite an verschiedenen Datenformaten erweitern (offen, kostenlos und maschinenlesbar) und folglich Daten zunehmend in nicht proprietären Formaten veröffentlichen. Um dieses Ziel zu erreichen, braucht es eine strukturelle Änderung des Formats für die tabellarischen Daten, die das BFS veröffentlicht. Insbesondere muss das Format XLSX/XLS möglichst rasch durch ein **nicht proprietäres Format** ersetzt werden, das wesentlich leichter von Maschinen gelesen und von anderen Anwendungen wiederverwendet werden kann (Austausch / Konvertierung).

2. Richtlinien für tabellarische Daten

Für tabellarische Daten empfehlen wir das Dateiformat **CSV** (Comma Separated Values), das häufig für den Austausch von Daten zwischen verschiedenen Anwendungen verwendet wird und **vermutlich die meistgenutzte Quelle für öffentliche Daten in einem offenen Format** darstellt. In der Regel kann eine Metadatendatei heruntergeladen werden oder ist zusammen mit der Datendatei in der beim Export erzeugten ZIP-Datei enthalten.

Eine CSV-Datei hat folgende Merkmale:
- Jede Variable hat eine eigene Spalte.
- Jede Beobachtung hat eine eigene Zeile.
- Jeder Wert hat eine eigene Zelle.

2.1 Eine Spalte pro Variable

Es darf weder Spaltenhierarchien noch Zellverbunde geben: Übergeordnete Kategorien werden in einer ersten Spalte geliefert, die untergeordneten Kategorien jeweils in einer weiteren Spalte.

2.2 Eine Zeile pro Beobachtung

Eine CSV-Datei darf keine leeren Zeilen enthalten. Wenn eine Beobachtung einer Erläuterung bedarf, ist diese auf eine der folgenden Arten mitzuliefern:
- in einer separaten Spalte (OBS_STATUS)
- in der Beschreibung des Datensatzes in der Metadatendatei

2.3 Eine Zelle pro Wert

Alle Felder in einer Spalte haben das gleiche Format. Die häufigsten Formate sind:
- alphanumerisch
- numerisch
- Datum

Bei numerischen Daten können die Werte frei gerundet werden, allerdings ist in den Metadaten darauf hinzuweisen, dass es sich um gerundete Daten handelt.

2.4 Regeln und Empfehlungen

Es ist wichtig, klare Regeln für die Produktion von CSV-Dateien festzulegen. Wenn man sich Empfehlungen zu diesem Dateiformat im Internet ansieht, zeigt sich, dass es diverse Dateiformate für unterschiedliche Anwendungen gibt. Das Ziel besteht letztendlich darin, offene Verwaltungsdaten bereitzustellen, die einfach zu nutzen sind.

```
"Périodes","Taux d'activité","Total","Suisses","Etrangers","Obs_Status Etrangers"
"2016-Q1","Taux d'activité brut (0-99 ans)","58.4888","56.6414","64.3373","A"
"2016-Q2","Taux d'activité brut (0-99 ans)","58.1806","56.4684","63.5852","A"
"2016-Q3","Taux d'activité brut (0-99 ans)","58.4418","56.8643","63.4159","A"
"2016-Q4","Taux d'activité brut (0-99 ans)","58.6242","56.9517","63.8775","A"
"2017-Q1","Taux d'activité brut (0-99 ans)","58.2404","56.5059","63.6746","A"
"2017-Q2","Taux d'activité brut (0-99 ans)","58.163","56.4049","63.6619","A"
"2017-Q3","Taux d'activité brut (0-99 ans)","58.3927","56.5487","64.1677","A"
"2017-Q4","Taux d'activité brut (0-99 ans)","58.3781","56.7083","63.5967","P"
"2016-Q1","Taux d'activité standardisé (15 ans et plus)","68.6253","66.2582","76.2143","A"
"2016-Q2","Taux d'activité standardisé (15 ans et plus)","68.2681","66.06","75.3262","A"
"2016-Q3","Taux d'activité standardisé (15 ans et plus)","68.5764","66.5247","75.1274","A"
"2016-Q4","Taux d'activité standardisé (15 ans et plus)","68.8036","66.6391","75.6886","A"
"2017-Q1","Taux d'activité standardisé (15 ans et plus)","68.3697","66.1359","75.456","A"
"2017-Q2","Taux d'activité standardisé (15 ans et plus)","68.285","66.0225","75.4504","A"
"2017-Q3","Taux d'activité standardisé (15 ans et plus)","68.5667","66.2005",,"L"
"2017-Q4","Taux d'activité standardisé (15 ans et plus)","68.5543","66.3967","75.3804","P"
```

Tabelle 1

Die CSV-Datei muss folgende Merkmale aufweisen:

- Die Zeichenkodierung ist «**UTF-8-BOM**»[1].
- Die Felder (Spalten) werden mit **einem Komma** getrennt.
- Sämtliche Felder (alphanumerisch, numerisch, Datum) stehen **in Anführungszeichen**.
- Die Enden von Zeilen in einer CSV-Datei sollten CRLF (U+000D U+000A) sein, können aber auch LF (U+000A) sein.
- Alphanumerische Felder dürfen keine Zeilenumbrüche enthalten.
- Die erste Zeile enthält die Spaltenüberschriften:
    - in Gross- und/oder Kleinbuchstaben;
    - mit Sonderzeichen (é, è, à, ô, ä, ö usw.);
    - mit oder ohne Leerzeichen (wenn es für den Bereich eine Ontologie gibt, ist diese für die Variablentitel zu verwenden).
- Die erste Zeile enthält nie eine Zeitreihe mit einer Zahl als Namen («2001», «2002», «2003»).
- Die erste Zeile umfasst weniger als 32 Zeichen (Empfehlung).

Die CSV-Datei kann eine oder mehrere Sprachen umfassen:

- Wenn sie einsprachig ist, aber in mehreren Sprachen verbreitet werden soll, ist eine Basissprache festzulegen und die Entsprechungen in den anderen Sprachen sind in der Metadatendatei anzugeben.

```
"year","week","date","recent","origin","value","obsvalue"
"2019","1","2019-01-02","adm","rall","oall","307923","A"
"2019","2","2019-01-09","adm","rall","oall","247856","A"
"2019","3","2019-01-16","adm","rall","oall","300043","A"
"2019","4","2019-01-23","adm","rall","oall","248896","A"
"2019","5","2019-01-30","adm","rall","oall","332925","A"
```

[1] Zeichen können von Programmen auf unterschiedliche Weise gespeichert werden. Damit maximale Kompatibilität mit anderen Programmen gewährleistet ist, muss unbedingt die UTF-8-BOM-Kodierung verwendet werden. Die Datei im Format CSV darf ausschliesslich Unicode-Zeichen enthalten, und zwar sowohl in den Spalten- und Zeilenüberschriften als auch in den einzelnen Werten. Andere Zeichensätze sind zu vermeiden, damit bei der Anzeige oder bei bestimmten Übertragungsvorgängen keine Probleme auftreten.

Tabelle 2

- Wenn sie mehrsprachig ist, muss der Variablentitel in den einzelnen Sprachen angegeben werden (wenn es für den Bereich eine Ontologie gibt, ist diese für die Variablentitel zu verwenden und am Ende der Spaltenüberschrift die Sprache «de, fr, it, en» anzugeben).

```
"Period","Erwerbsquoten","Taux d'activité","Total","Schweizer","Suisses"
"2020-01","Bruttoerwerbsquoten (0-99 Jahre)","Taux d'activité brut (0-99 ans)","58.5","56.6","56.6"
"2020-02","Bruttoerwerbsquoten (0-99 Jahre)","Taux d'activité brut (0-99 ans)","58.2","56.5","56.5"
"2020-03","Bruttoerwerbsquoten (0-99 Jahre)","Taux d'activité brut (0-99 ans)","58.4","56.9","56.9"
"2020-04","Bruttoerwerbsquoten (0-99 Jahre)","Taux d'activité brut (0-99 ans)","58.6","56.9","56.9"
```

Tabelle 3

Der Datentyp (datatype) jeder Spalte (Variable) muss in der Metadatendatei klar angegeben werden:

- alphanumerisch (String)
- numerisch (Integer / Float)
- Datum (nach ISO 8601)

Numerische Daten enthalten **kein Tausendertrennzeichen** in den Zahlenwerten. Wenn kein numerischer Wert gemessen wurde, bleibt das Feld leer. Eine CSV-Datei kann mehrere Variablen des Typs «numerisch» enthalten. Ein numerisches Feld muss entweder einen Zahlenwert enthalten oder leer bleiben: Im zweiten Fall ist darauf zu achten, dass der Status des leeren Felds in der folgenden Spalte beschrieben und in der Metadatendatei erklärt wird. Bei relativen Daten (Float) **ist das Dezimaltrennzeichen immer ein Punkt**.

Spalten des Typs «alphanumerisch» dürfen keine leeren Felder enthalten!

```
"Périodes","Taux d'activité","Suisses","Etrangers","Obs_Status Etrangers"
"2016-Q1","Taux d'activité standardisé (15 ans et plus)","68.6253","66.2582","A"
"2016-Q2","Taux d'activité standardisé (15 ans et plus)","68.2681",,"L"
"2016-Q3","Taux d'activité standardisé (15 ans et plus)","68.5764","66.5247","A"
```

Tabelle 4

Für Felder des Typs «Datum» sind nebst dem Format gemäss ISO-8601 auch andere Formate zulässig, sofern sie in der Metadatendatei beschrieben werden:

- Halbjahresdaten: 2020-S1 (1. Halbjahr 2020) und 2020-S2 (2. Halbjahr 2020)
- Quartalsdaten: 2020-Q1 (1. Quartal 2020) bis 2020-Q4 (4. Quartal 2020)
- Monatsdaten: 2020-M01 (Januar 2020) bis 2020-M12 (Dezember 2020)
- Wochendaten: 2020-W36 (Woche 36, 2020)
- Daten für ein Zeitintervall: 2016/2018

Es gilt zu beachten, dass Kommentare mit dem Zeichen # oder Bemerkungen in jeglicher Form in CSV-Dateien nicht zulässig sind.

2.4.1 Weitere Empfehlungen

Die Angaben zur Einheit für ein numerisches Feld müssen in einer separaten Spalte erfasst werden.

```
"Catégorie d'émissions","Année","Unité de mesure","Emissions","Obs_status"
"Total","1990","Milliers de tonnes","57237.468","A"
"Total","1990","Tonnes par habitant","8.479","A"
"Total","1991","Milliers de tonnes","59145.409","A"
"Total","1991","Tonnes par habitant","8.643","A"
"Total","1992","Milliers de tonnes","58517.892","A"
"Total","1992","Tonnes par habitant","8.471","A"
```

Tabelle 5

Die Spaltentitel (Variablen) müssen so sprechend wie möglich sein. Wenn dies nicht möglich ist, muss die Variablenbezeichnung in der Beschreibung (Metadatendatei) des entsprechenden Datensatzes erläutert werden.

Das Fehlen eines numerischen Werts (unbekannt, keine Angabe, Datenschutz usw.) muss in einer zusätzlichen Variable begründet werden (als OBS_STATUS).

```
"Périodes","Taux d'activité","Total","Suisses","Etrangers","Obs_Status Etrangers"
"2017-Q3","Taux d'activité brut (0-99 ans)","58.3927","56.5487","64.1677","A"
"2017-Q4","Taux d'activité brut (0-99 ans)","58.3781","56.7083","63.5967","P"
"2017-Q3","Taux d'activité standardisé (15 ans et plus)","68.5667","66.2005",,"L"
"2017-Q4","Taux d'activité standardisé (15 ans et plus)","68.5543","66.3967","75.3804","P"
```

Tabelle 6

Der wichtigste Unterschied zum «maschinenlesbaren» Format, das wir vor zwei Jahren vorgestellt haben, besteht darin, dass mit CSV eine unbegrenzte Anzahl Spalten (Variablen) mit numerischen Werten bereitgestellt werden kann.

3. **Metadatendatei**

Jede Diffusion im Format CSV erfordert ein dazugehöriges Dokument, das den Inhalt der CSV-Datei beschreibt: die Metadatendatei. Sie muss in einem offenen Format vorliegen, z.B. TXT oder ODS.

Die Metadatendatei muss zwingend folgende Informationen enthalten:

- ein Titel, der die statistischen Daten in der CSV-Datei beschreibt
- Hinweise und Bemerkungen zu besonderen Aspekten oder Merkmalen der Daten
- Angaben zur Herkunft der Daten (Institution, Quelle/Erhebung) und Copyright (mit Jahresangabe)
- nützliche Zusatzinformationen (Telefonnummern, E-Mail-Adressen)
- eine Beschreibung jeder einzelnen Variable in der CSV-Datei (falls der Titel nicht sprechend ist) und deren Datentyp (datatype)
- Übergangsschlüssel mit den Texten in den verschiedenen Sprachen, falls für bestimmte Variablen Codes verwendet werden.

4. **Schlussfolgerungen**

Es gibt verschiede Modelle für tabellarische Daten. Die grosse Vielfalt an CSV-basierten Dateiformaten ist unter anderem auf die **bisher unzureichende Standardisierung des Formats (keine offizielle Norm oder Spezifizierung)** zurückzuführen. So werden beispielsweise unterschiedliche Trennzeichen oder Tabellen mit festem Format, mehrere Tabellen innerhalb einer

Datei oder Tabellen mit Metadaten-Zeilen im Tabellenkopf verwendet. Dieses Dokument beschreibt einige Regeln für eine sichere Nutzung dieses Formats.

Für Daten mit komplexen Strukturen oder für die Datenübertragung zwischen verschiedenen Programmen und Systemen empfehlen wir die Formate JSON und XML.

Pierre-Alain Baeriswyl / Edy Juillerat
Data@bfs.admin.ch

Anhang: Liste und Erläuterungen der in der Spalte «OBS_STATUS» verwendeten Symbole

Die Liste umfasst einerseits für das BFS-Portal bereits verwendete Symbole (hauptsächlich im Statistischen Jahrbuch) und andererseits bestehende Elemente aus dem SDMX-Bereich.

Code	Name_fr	Name_de
A	Valeur normale	Normaler Wert
B	Rupture de la série chronologique	Zeitreihenbruch
D	La définition diffère	Definition unterscheidet sich
E	Valeur estimée	Geschätzter Wert
F	Valeur prévisionnelle	Prognostizierter Wert
G	Valeur expérimentale	Experimenteller Wert
I	Valeur imputée par une agence réceptrice	Von einer Empfangsstelle unterstellter Wert
K	Données incluses dans une autre catégorie	Daten, die in einer anderen Kategorie enthalten sind
W	Inclut les données d'une autre catégorie	Enthält Daten aus einer anderen Kategorie
O	Valeur manquante	Fehlender Wert
M	Valeur manquante ; les données ne peuvent pas exister	Fehlender Wert; Daten können nicht existieren
P	Valeur provisoire	Provisorischer Wert
S	Grève et autres événements spéciaux	Streik und andere besondere Ereignisse
L	Valeur manquante ; les données existent mais n'ont pas été collectées.	Fehlender Wert; Daten sind vorhanden, wurden aber nicht erhoben
H	Valeur manquante ; vacances ou week-end	Fehlender Wert; Feiertag oder Wochenende
Q	Valeur manquante ; supprimée	Fehlender Wert; unterdrückt
J	Dérogation	Ausnahmeregelung
N	Non significatif	Nicht signifikant
U	Faible fiabilité	Geringe Zuverlässigkeit
V	Valeur non validée	Nicht validierter Wert
R	Valeurs révisées	Revidierte Werte

Explications :

Code	Desc_en
A	To be used as default value if no value is provided or when no special coded qualification is assumed. Usually, it can be assumed that the source agency assigns sufficient confidence to the provided observation and/or the value is not expected to be dramatically revised.
B	Observations are characterised as such when different content exists or a different methodology has been applied to this observation as compared with the preceding one (the one given for the previous period).
D	Used to indicate slight deviations from the established methodology (footnote-type information); these divergences do not imply a break in time series.
E	Observation obtained through an estimation methodology (e.g. to produce back-casts) or based on the use of a limited amount of data or ad hoc sampling and through additional calculations (e.g. to produce a value at an early stage of the production stage while not all data are available). It may also be used in case of experimental data (e.g. in the context of a pilot ahead of a full scale production process) or in case of data of (anticipated/assessed) low quality. If needed, additional information can be provided through free text using the COMMENT_OBS attribute at the observation level or at a higher level . This code is to be used when the estimation is done by a sender agency. When the imputation is carried out by a receiver agency in order to replace or fill gaps in reported data series, the flag to use is I “Value imputed by a receiving agency”.
F	Value deemed to assess the magnitude which a quantity will assume at some future point of time (as distinct from "estimated value" which attempts to assess the magnitude of an already existent quantity).
G	Data collected on the basis of definitions or (alternative) collection methods under development. Data not of guaranteed quality as normally expected from provider.
I	Observation imputed by a receiving agency to replace or fill gaps in reported data series. This code is intended to cover all cases where a receiving agency publishes data about a sending

	agency that do not come from an official source in the sender agency's reporting framework. When the estimation is done by the sender agency, the flag to use is E “Estimated value”.
K	This code is used when data for a given category are missing and are included in another category, sub-total or total. Generally where code “K” is used there should be a corresponding code "W - Includes data from another category" assigned to the over-covered category. Implementers and data reporters should use the COMMENT_OBS observation-level attribute to specify under which category the data are included.
W	This code is used when data include another category, or go beyond the scope of the data collection and are therefore over-covered. Generally, where code "W" is used there should be a corresponding code "K - Data included in another category" assigned to the category which is under-covered. Implementers and data reporters should use the COMMENT_OBS observation-level attribute to specify which additional data are included.
O	This code is to be used when no breakdown is made between the reasons why data are missing. Data can be missing due to many reasons: data cannot exist, data exist but are not collected (e.g. because they are below a certain threshold or subject to a derogation clause), data are unreliable, etc.
M	Used to denote empty cells resulting from the impossibility to collect a statistical value (e.g. a particular education level or type of institution may be not applicable to a given country's education system).
P	An observation is characterised as "provisional" when the source agency – while it bases its calculations on its standard production methodology – considers that the data, almost certainly, are expected to be revised.
S	Special circumstances (e.g. strike) affecting the observation or causing a missing value.
L	Used, for example, when some data are not reported/disseminated because they are below a certain threshold.
H	Used in some daily data flows.
Q	Used, for example, when data are suppressed due to statistical confidentiality considerations.
J	Clause in an agreement (e.g. legal act, gentlemen’s agreement) stating that some provisions in the agreement are not to be implemented by designated parties; these derogations may affect the observation or cause a missing value. In general, derogations are limited in time.
N	Used to indicate a value which is not a "real" zero (e.g. a result of 0.0004 rounded to zero).
U	This indicates existing observations, but for which the user should also be aware of the low quality assigned.
V	Observation as received from the respondent without further evaluation of data quality.
R	Revised Data.

Referenzen im Internet

W3C-Empfehlungen für das Format CSV:
CSV on the Web: A Primer

Anforderungen für die Datenvorbereitung:
Factsheet Linked Data Plattform Bund

Empfehlungen für eine hohe Daten- und Metadatenqualität:
Leitfaden für qualitativ Hochwertige

ISO-8601-Standard für Datums- und Zeitformate:
https://www.iso.org/iso-8601-date-and-time-format.html
https://en.wikipedia.org/wiki/ISO_8601

Common Format and MIME Type for CSV Files:
RFC 4180 officielle

Modelle für tabellarische Daten und Metadaten:
Model for Tabular Data and Metadata on the Web
https://github.com/BurntSushi/rust-csv/issues/46

SDMX-CSV field guide:
https://github.com/sdmx-twg/sdmx-csv/blob/master/data-message/docs/sdmx-csv-field-guide.md