Tabellenkalkulation: Situationsanalyse und Perspektiven für den Katalog archivischer Dateiformate KaD

Lambert Kansy, Staatsarchiv Basel-Stadt,
unter Mitarbeit der Preservation-Planning-Expertengruppe PPEG der KOST
September 2021

1 Einleitung

1.1 Vorbemerkung

Die vorliegende Studie entstand im Rahmen der Erarbeitung von Version 6.2 des *Katalogs archivischer Dateiformate KaD*[1] der KOST. Die Überarbeitung der Empfehlungen zu Tabellenkalkulationsformaten bildete einen Schwerpunkt dieser Version. Die Studie gibt den Erkenntnisstand von September 2021 wieder; Version 6.2 des KaD wurde im Dezember 2021 publiziert.

1.2 Ausgangslage und Fragestellung

Tabellenkalkulationssoftware stellt numerische und alphanumerische Daten dar und erlaubt insbesondere deren Verarbeitung mittels Funktionen. Seit der Version 5.1 von 2017 empfiehlt der KaD folgende Vorgehensweise bei der Archivierung von Tabellenkalkulationsdateien:

«Als provisorische Lösung sollen Tabellenkalkulationsdaten im Originalformat archiviert werden, also in der Regel in XLS, ODF oder OOXML. Die Alternativen dazu, nämlich die Archivierung in Datenbank-Archivformaten oder die Konvertierung in PDF/A, können nicht als Best Practice gelten und müssen kritisch beurteilt werden.»[2]

Es stellt sich die Frage, ob es seit 2017 neue Erkenntnisse gibt, die eine Veränderung der Empfehlung bewirken. Die grundsätzlich zu berücksichtigenden Aspekte werden hierbei als unverändert geltend vorausgesetzt:

a. Funktionalität (Aspekt der Kalkulation)
b. Tabellennatur
c. optische Erscheinung

[1] https://kost-ceco.ch/cms/dateiformate.html
[2] https://kost-ceco.ch/wiki/whelp/KaD_cmsv5.1/tabellenkalkulation-einleitung.html

2 Abgrenzung

Die Ausführungen beziehen keine Cloud-basierten Formate wie Google Docs ein. Ebenso werden keine vertieften Abklärungen zur Interoperabilität zwischen ODF und OOXML und zu den Risiken bei der Konvertierung der jeweiligen Vorläuferformate (XLS resp. SXC) in die aktuellen Formate durchgeführt. Auch wird nicht abgeklärt, welches der beiden aktuellen Formate (XLSX und ODS) als Zielformat für die Konvertierung aus anderen, obsoleten Dateiformaten unter welchen Umständen besser geeignet ist.

3 Best Practice

Aufgrund dieser Fragen wurden die aktuellen Empfehlungen zum Umgang mit Tabellenkalkulationsdaten folgender Archive geprüft[3]:

- Nationaal Archief (NL): Handreiking voorkeursformaten Nationaal Archief, **2016** https://www.nationaalarchief.nl/archiveren/kennisbank/handreiking-voorkeursformaten-nationaal-archief
- SLUB Dresden, Langzeitarchivfähige Dateiformate, Version 1.3.4, **2021-02-22** https://slubarchiv.slub-dresden.de/fileadmin/groups/slubsite/slubarchiv/SLUBArchiv_langzeitarchivfaehige_Dateiformate_v1.3.4.pdf
- ETH-Bibliothek, Archivtaugliche Dateiformate, **s. d.** https://documentation.library.ethz.ch/display/DD/Archivtaugliche+Dateiformate
- Library and Archives Canada, Guidelines on File Formats for Transferring Information Resources of Enduring Value, **2014-10-01** https://www.bac-lac.gc.ca/eng/services/government-information-resources/guidelines/Documents/file-formats-irev.pdf
- Archaeology Data Service, GUIDELINES FOR DEPOSITORS: Downloads, Version 4.1, **April 2021** https://archaeologydataservice.ac.uk/advice/Downloads.xhtml
- Archaeology Data Service / Digital Antiquity, Guides to Good Practice: Databases and Spreadsheets: A Guide to Good Practice, **2013** https://guides.archaeologydataservice.ac.uk/g2gp/DbSht_1
- Florida Digital Archive, Recommended Data Formats for Preservation Purposes in the Florida Digital Archive, **2013** https://libraries.flvc.org/documents/181844/502298/Recommended+Data+Formats/0b25496f-33ac-4f56-9550-12c34f3d5d7c
- UK Data Archive, Recommended formats, **s. d.** https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats
- NARA, Transfer Guidance Format, Structured Data Formats, **s. d.** https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html#structuredata
- The National Archives, File formats for transfer, **s. d.** https://www.nationalarchives.gov.uk/information-management/manage-information/digital-records-transfer/file-formats-transfer/
- Library of Congress, Recommended Formats Statement, VI. Datasets, **s. d**. (https://www.loc.gov/preservation/resources/rfs/data.html

[3] Wenn nicht anders vermerkt, wurden alle aufgeführten Websites und Links in dieser Studie am 30.06.2021 letztmals aufgerufen.

- Schweizerisches Bundesarchiv, Archivtaugliche Dateiformate, Version 2020/04 **April 2020**
  https://www.bar.admin.ch/dam/bar/de/dokumente/konzepte_und_weisungen/archivtaugliche_dateiformate.1.pdf.download.pdf/archivtaugliche_dateiformate.pdf

Das niederländische Nationalarchiv bezeichnet als bevorzugte Formate für die Archivierung von Tabellenkalkulationsdaten **ODS**, CSV und PDF/A, während XLS und **XLSX** akzeptierte Formate sind.

Die SLUB Dresden bezeichnet die **OOXML**-Formate als archivtaugliche Formate, sofern sie mit MS Office 2013 und neuer erstellt worden sind. Die **ODF**-Formate werden uneingeschränkt als archivtaugliche Formate bezeichnet.

Die ETH-Bibliothek bezeichnet als geeignete Formate für Nutzung über mehr als 10 Jahre lediglich CSV und TAB-delimited, während OOXML (**XLSX**) für einen Nutzungszeitraum von bis zu 10 Jahren akzeptiert wird. **XLS** hingegen wird nicht mehr als archivtauglich angesehen. **ODS** wird nicht genannt.

Library and Archives Canada präferieren ASCII-Text und CSV, akzeptieren aber auch DBF, EBCDIC, **OOXML**, **ODS** und XLS. Es werden klare Anforderungen an die deskriptiven Metadaten formuliert.

Der Archaeology Data Service bezeichnet CSV als bevorzugtes Format, während **XLSX** und **ODS** als Format gewählt werden, wenn Funktionalitäten erhalten werden sollen. Ältere Formate von MS-Office und ODF sollen nach XLSX und ODS konvertiert werden. Es werden Vorgaben für die Beschreibung von Tabellenkalkulationsdaten gemacht. In der ausführlichen Erörterung der Erhaltung von Spreadsheet-Daten werden **XLSX** und **ODS** als geeignete Formate für eine dauerhafte Aufbewahrung bezeichnet, sofern gewissen Eigenschaften der Formate Rechnung getragen wird und adäquate Massnahmen getroffen werden.

Das Florida Digital Archive teilt CSV, TXT und SQL DDL der Gruppe von Formaten zu mit „High Confidence Level“, während in die Gruppe mit „Medium Confidence Level“ die Formate DBF, **ODS** resp. SXC und **XLSX** eingeteilt werden. XLS wird der Gruppe „Low Confidence Level“ zugewiesen.

Das UK Data Archive bezeichnet CSV und TAB als akzeptable Formate für Sharing, Reuse and Preservation, während TXT sowie XLS und **XSLX** neben **ODS,** aber auch MDB, als akzeptable Formate für den Datenerhalt bezeichnet werden.

Die Empfehlungen der bisher genannten Archive sind seit 2017 kaum verändert worden. Neu betrachtet wurden die Empfehlungen von NARA, The National Archives, Library of Congress und des Schweizerischen Bundesarchivs.

Die NARA empfiehlt für strukturierte Daten als «Preferred Formats» CSV, ODS, ASCII Textm JSON und XML. Als «Acceptable Formats» gelten Microsoft Excel Office Open XML und Microsoft Excel 97 Binary File Format (XLS)

The National Archives listet neben CSV OpenDocument Spreadsheet, dessen Vorgängerformat OpenOffice Calc, Lotus 1-2-3 Worksheet und Microsoft Excel sowie Microsoft Works in nahezu allen Versionen auf.

Die Library of Congress bezeichnet als «Preferred» plattformunabhängige, zeichenbasierte Formate und führt dazu aus:

«Preferred formats include well-developed, widely adopted, de facto marketplace standards, e.g.

a. Formats using well known schemas with public validation tool available
b. Line-oriented, e.g. TSV, CSV, fixed-width
c. Platform-independent open formats, e.g. .db, .db3»

Weiterhin zählen zu den bevorzugten Dateiformaten auch «Any proprietary format that is a de facto standard for a profession or supported by multiple tools (e.g. Excel .xls or .xlsx, Shapefile)». Die verwendeten Zeichensätze sind in absteigender Reihenfolge UTF-8 und UTF-16 (with BOM), US-ASCII oder ISO 8859-1 und «Other named encoding»

Als «Acceptable» gelten Formate, die folgende Eigenschaften aufweisen: «Non-proprietary, publicly documented formats endorsed as standards by a professional community or government agency, e.g. CDF, HDF» und «Text-based data formats with available schema»

Das Schweizerische Bundesarchiv empfiehlt CSV als Dateiformat für die Archivierung von Tabellen.

4 OOXML und ODF

4.1 OOXML

Office Open XML (OOXML) wurde zuerst 2006 als ECMA-Standard ECMA-376. standardisiert. 2008 wurde es zudem als ISO 29500 mit vier Teilen standardisiert, allerdings in zwei Versionen, «Transitional» und «Strict». ECMA-376 ist nicht vollständig kompatibel mit ISO 29500. Die aktuelle Version ist ISO/IEC 29500:2016

Die Version «Strict» von ISO 29500 wird erst mit Microsoft Office 2013 voll unterstützt. Microsoft Office 2007 unterstützt nur ECMA-376 und Microsoft Office 2010 kann zwar «Strict» und «Transitional» Dokumente lesen, jedoch Dokumente nur in der «Transitional»-Version speichern.

Siehe zu OOXML/XLSX:
- Library of Congress, XLSX Transitional (Office Open XML), ISO 29500:2008-2016, ECMA-376, Editions 1-5 https://www.loc.gov/preservation/digital/formats/fdd/fdd000398.shtml
- Wikipedia, Office Open XML (de) https://de.wikipedia.org/wiki/Office_Open_XML
- Wikipedia, Office Open XML (en) https://en.wikipedia.org/wiki/Office_Open_XML

4.2 ODF

Das OpenDocument-Format wurde 2005 von der Organization for the Advancement of Structured Information Standards (OASIS) als Standard publiziert. Das Format Open Document Format for Office Applications (OpenDocument) v1.0 wurde 2006 als ISO 26300 publiziert. Die Version 1.2 wurde in drei Teilen 2015 als ISO/IEC 26300:2015 publiziert; dies ist die aktuelle Version als ISO-Norm. ODF 1.3 wurde 2019 von der OASIS publiziert.

Siehe zu ODF/ODS:
- Library of Congress, OpenDocument Spreadsheet Document Format (ODS), Version 1.2, ISO 26300:2015 https://www.loc.gov/preservation/digital/formats/fdd/fdd000439.shtml
- Wikipedia OpenDocument (de) https://de.wikipedia.org/wiki/OpenDocument
- Wikipedia OpenDocument (en) https://en.wikipedia.org/wiki/OpenDocument

4.3 Interoperabilität von OOXML und ODF

Es ist nicht ohne weiteres möglich, Tabellenkalulationsdateien verlustfrei in Bezug auf Layout und Funktionen von ODF nach OOXML oder umgekehrt zu konvertieren. Für die detaillierte Argumentation verweisen wir auf die Untersuchungen der Library of Congress und des Fraunhofer-Instituts:

- Library of Congress, XLSX Transitional (Office Open XML), ISO 29500:2008-2016, ECMA-376, Editions 1-5, Notes/General : Conversion between XLSX and ODS (https://www.loc.gov/preservation/digital/formats/fdd/fdd000398.shtml [aufgerufen 01.07.2021])
- Fraunhofer-Institut für Offene Kommunikationssysteme, FOKUS (Hg.), Interoperabilität von Dokumentenformaten: Open Document Format und Office Open XML; white paper, Stuttgart 2009

5 Fazit

Zusammenfassend lässt sich sagen, dass die Verwendung von **ODF** und **OOXML** als archivische Dateiformate für den Bereich der Tabellenkalkulation nach wie vor angeraten ist. Es ist kein neues Format in Sicht, das in Bezug auf Verbreitung und Standardisierung mit diesen beiden Formaten mithalten kann. Gleichwohl ist diese Empfehlung mit zwei Vorbehalten zu versehen.

Zum einen sind bei beiden Formaten nicht alle Versionen gleichermassen sinnvoll einsetzbar. Bei OOXML ist ISO 29500 Version «Strict» als empfohlene Format-Version zu sehen. Dies bedeutet etwa beim Einsatz der Microsoft Office-Pakete, dass Microsoft Office 2013 oder jünger eingesetzt werden muss, damit diese Version auch erzeugt und gespeichert werden kann. Beim Einsatz anderer Office-Werkzeuge muss dieses gezielt abgeklärt werden. Bei ODF ist die Version 1.2 anzustreben. Diese wurde als neueste Version als ISO-Norm standardisiert, während Version 1.3 nur von der OASIS genormt wurde.

Zum anderen gilt die Empfehlung, Dateien in älteren Versionen auf die empfohlenen Versionen zu aktualisieren, nach wie vor; insbesondere für das binäre Dateiformat XLS sowie die Vorgänger-Formate von ODF. Es sind aber nur wenige Informationen greifbar, ob und welche Veränderungen und Informationsverluste damit einhergehen respektive unter welchen Rahmenbedingungen solches erfolgen kann.

Eine vertiefte Auseinandersetzung mit diesen beiden Formaten erscheint nach wie vor angebracht. Auffällig ist, dass es nur wenig aktuelle, spezifisch auf Tabellenkalkulation ausgerichtete Literatur gibt. Die bis und mit Version 6.1 einzige Referenz im KaD stammt von 2003. 2021 ist eine neue Studie erschienen, die den aktuellen Stand zusammenfasst und auf weitere Literatur verweist: Artefactual Systems, Digital Preservation Coalition (eds.), Preserving Spreadsheets. DPC Technology Watch Guidance Notes, Data Types Series. 2021. http://doi.org/10.7207/twgn21-09 [aufgerufen 08.09.2021]. Hinzu kommen die Informationen der Library of Congress, Sustainability of Digital Formats: Planning for Library of Congress Collections, https://www.loc.gov/preservation/digital/formats/fdd/dataset_fdd.shtml [aufgerufen 09.09.2021]