Binary Separated Value

B. Bean, N. Mezher, J. Summers & D. Toomey
University of Massachusetts Lowell — Software Engineering II
Prof. James Daly

April 2020

Motivation

The use of databases as a means for storing information has become ubiquitous in every field concerning data. One of the most common methods for analyzing sets of data from databases is to export it to a `.csv` (comma separated value) file format in order to manipulate the data via a spreadsheet program or language libraries. But despite all its draws, the `.csv` format has some significant drawbacks as well. The format is too bulky and inefficient for many applications, and it relies on a comma delimiter to separate data which can be problematic [1]. Our improvements upon the `.csv` format allow for users and programs to more efficiently store and utilize complex data. The end goal is to expedite communication between programs and disparate systems.

This document first outlines a new file format termed Binary Separated Value (BSVX) and with the extension `.bsvx`. This is not to be confused with the `.bsv` file format, which is a BASIC BSave Graphics file. The tailing `x` was chosen for convenience as it makes the name of our format, `.bsvx`, wholly unique. This format of data is delimited with byte markers which begin each field telling the library what kind of data will be in the field and how long it will be. Using byte markers, instead of plaintext character delimiters, solves a key issue with the `.csv` format—strings including commas do not prematurely end a field. The Binary Separated Value format is processed through a proprietary Python library.

One drawback of this style of implementation is the inability to parse and edit `.bsvx` files through a text editor. However, this issue is remedied through the BSVX LibreOffice Calc Extension. LibreOffice is a free to use, open-source file editing platform similar to Microsoft Office. Calc is a program provided in the LibreOffice suite, and provides similar functionality to Microsoft's Excel program [4]. The BSVX LibreOffice Calc Extension gives Calc users the ability to read data from `.bsvx` files and export their spreadsheets to `.bsvx` files. Users also have the ability to import, then convert `.csv` files into `.bsvx` files through the BSVX LibreOffice Calc Extension.

BSVX File Format Specification

Each `.bsvx` file contains a series of rows of headers or records. Each row begins with a byte marker denoting the type (i.e. header or record) and the number of fields within that row. Following the first byte marker of a row is a series of fields, each made up of two parts: a byte marker denoting the type and size of the data stored within the field, and the data itself. Some initial markers indicate that the size of the data is given in subsequent bytes. Once the length n is determined, those n bytes can be interpreted to match the field byte marker. Each row does not have to be the same length, the data can be jagged and parsers read as much data as is denoted by the first byte marker of each row.

At any time, the parser knows how many bytes it needs to read. There is never an instance where the parser needs to read bytes until it sees a particular character (as opposed to `.csv` or `.tsv` parsers, which look for commas or tabs respectively). Strings need neither end marks nor escape characters, and are stored in the UTF-8 format. The byte marker for strings denotes the number of bytes read, not the number of characters of the string itself. All numbers are stored in little endian order.

An abstract example of a `.bsvx` file row (header or record) looks like this:

3	3	FOO	2	1000	4	25.345
Field	str		int		Float	

Table 1: An example of a `.bsvx` file row.

The same example of a `.bsvx` file row (header or record) represented in hexadecimal:

0xAB	0x03	0x464F4F	0x91	0x01F4	0x9C	0x400395851EB851EB851EB851EB851EB8
Record	str		int		Float	

Table 2: An example of a `.bsvx` file row in hex values.

The following table will be used to implement each type of supported data in its own class and denote the bit range each field will be denoted by, ranging from 0 to 255. The first column gives the parser crucial context: what type of data will follow the byte marker, and further, which *variant* on that type it will be. For example, the small integer (int) type is represented by numbers in the range 136-143. A 2-byte small int is indicated by 138, 139 is for a 3 byte int, 140 denotes a 4 byte int, etc. The second column illustrates how the range of values for a given type is affected by the magnitude of its offset i.e. for a small int, the second column entry is $136 + [0, 7]$. The third column establishes the types of data that are supported and the fourth provides a brief description of each.

Range	Form	Name	Description
0		Blank	Possible implementation: NULL or 'empty string'
1-127	1-127	Short str	UTF-8 Encoded string of byte length 1-127
128-135	$128 + [0,7]$	Long str	1-8 bytes giving the length of a str, followed by said str
136-143	$136 + [0,7]$	Short int	An integer in the range of 0-7 bytes
144-151	$144 + [0,7]$	Long int	A zig-zag encoded integer using 1-8 bytes
152-159	$152 + [0,7]$	Float	IEEE-754 format float: 0 = half precision, 1 = single, 2 = double, 3 = triple
160-167	$160 + [0,7]$	Blob	1-8 bytes giving the length of binary data in bytes, followed by said data
168-183	$168 + [0,15]$	Header	Beginning of header with 0-15 fields
184-191	$184 + [0,7]$	Long header	1-8 bytes giving the number of fields in the header
192-207	$192 + [0,15]$	Record	Beginning of record with 0-15 fields
208-215	$208 + [0,7]$	Long record	1-8 bytes giving the number of fields in the record
216-255		Reserved	For future use

Table 3: Data types supported by the BSVX file type specification.

Deliverables

The proposed deliverables for this new format include a LibreOffice Calc extension to read from and write to `.bsvx` files, and a Python library for doing the same. This extension should be capable of converting between `.csv` and `.bsvx` without loss or adulteration of the information stored in the files. Similar libraries for languages such as Java, C++, or JavaScript are left as stretch goals.

The BSVX Calc Extension will allow spreadsheets to be saved to and read from `.bsvx` files. To illustrate the top-most point for user interaction with the BSVX Calc Extension, there is included a series of figures below. Figure 1 displays the default toolbar packaged with LibreOffice Calc. Figure 2 contrasts the differences between the default toolbar and the toolbar with the BSVX Calc Extension enabled. It can be seen that only two features are added, in the form of two buttons. The proposed left button allows for importing, reading from, a `.bsvx` file and the proposed right button for exporting, saving to, a `.bsvx` file. Finally, Figure 3 provides a glance as to how the toolbar will look with the BSVX Calc Extension enabled.
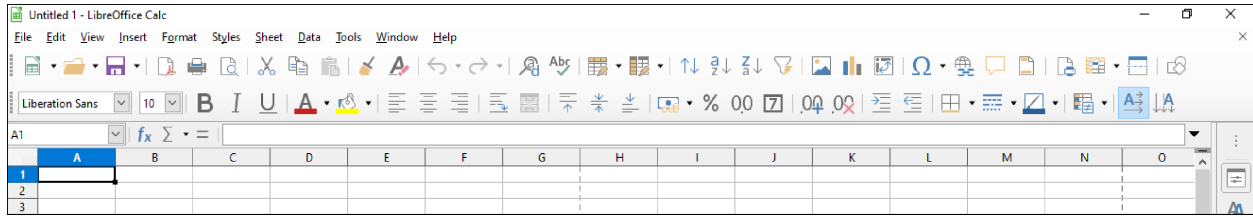
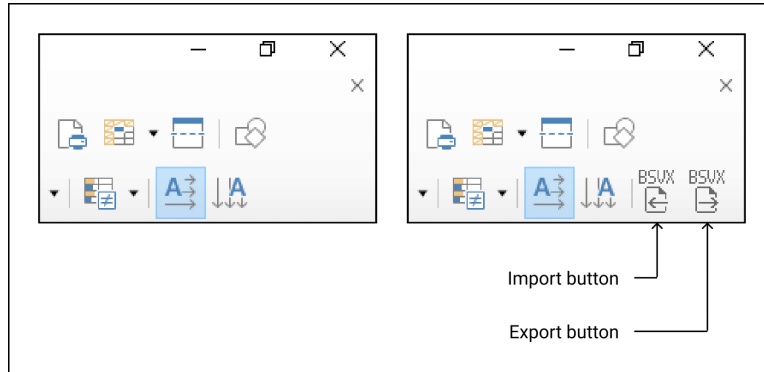Figure 1: LibreOffice Calc's toolbar.

Figure 2: The BSVX Calc Extension will provide two additional buttons for importing and exporting.
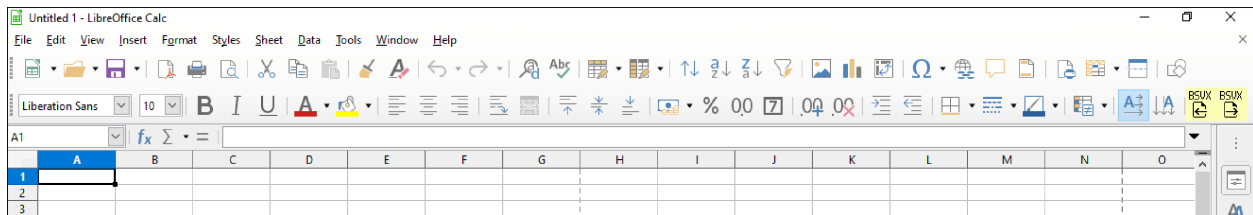

Figure 3: A mockup of LibreOffice Calc's toolbar with the BSVX Calc Extension enabled.

To talk more about how the BSVX Calc Extension works behind the scenes, it is first necessary to speak about our proposed Python library—bsvxpy. As a generality, our Python library will be similar to the .csv library. A writer function will be passed a series of fields representing a header row. Subsequent binary values are decoded based on the corresponding type casts provided by the header. The library will then extract each of the fields from the dictionary object and output them to the Calc spreadsheet in sequential order. The library will also process nested data structures within the .bsvx file allowing for the recursive encoding and decoding of further dictionary objects.
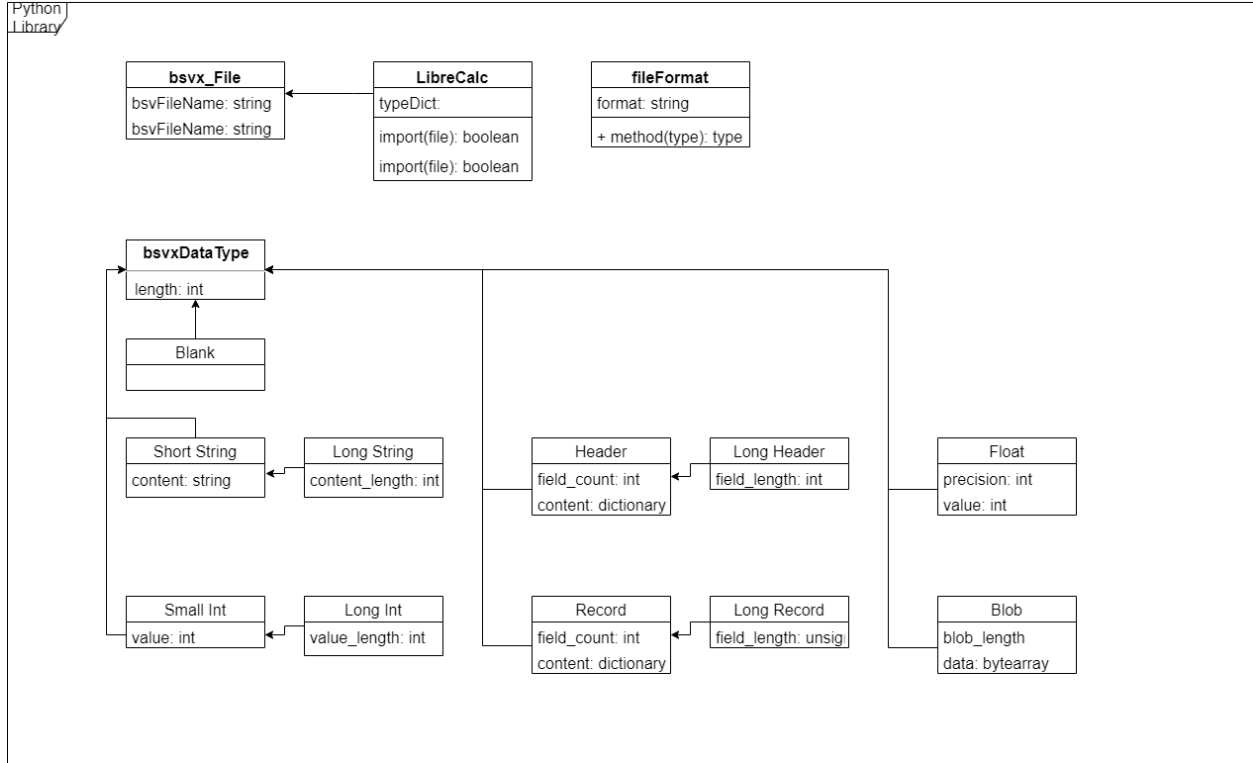
Figure 4: UML diagram outlining the classes and statstructures used in bsvxpy

With an understanding of how the `bsvxpy` Python library functions, we can return to an overview of the BSVX Calc Extension. First off, the LibreOffice Calc project allows developers to create extensions using Python, which lets us extend LibreOffice Calc's functionality to include `.bsvx` file support using our `bsvxpy` Python library. We will also be using the `uno` Python library, as it is necessary for any LibreOffice Calc extension development.

A novel problem with strongly typed encoding is the inability to handle undefined data types such as graphs or algorithms that are often used in spreadsheet applications. The `.bsvx` file format manages unknown data types with a catch-all data type: Blob, which acts as a polymorphic object with which to store raw binary data exported from the spreadsheet or csv file. By relying on the raw binary for unknown data types, `bsvxpy` can accommodate proprietary encoding s associated with any 3rd part application; however, this introduces the risk of `.bsvx` files being limited to one spreadsheet application when using Blob data types, as the 3rd party application may use internal encoding which are unknown to other programs.

To export data to a `.bsvx` file, the BSVX Calc Extension will call functions from the `uno` Python library to read cell data from LibreOffice Calc. The `bsvxpy` Python library will then be used to convert that data into binary separated values. Once the data is appropriately converted, it will be written to a file—as named by the user—with the file extension `.bsvx`. Figure 5 depicts the data flow for the exporting feature.
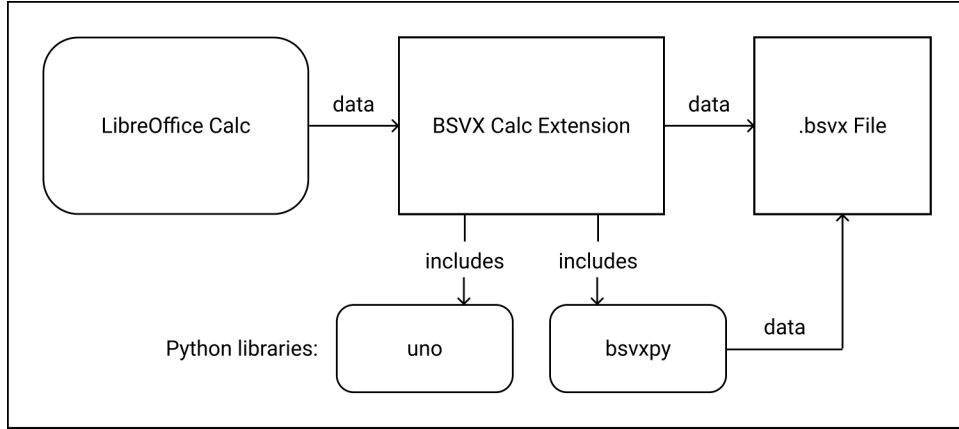
Figure 5: Scope of API and library calls for the BSVX Calc Extension in exporting data to a `.bsvx` file.

The importing feature works the same way, but in reverse. The user selects a `.bsvx` file to import data from, and the BSVX Calc Extension reads data from that file, using `bsvxpy` and `uno` to translate it from binary separated value data to cell data that LibreOffice Calc can read. The data flow for the importing feature is represented by Figure 6.
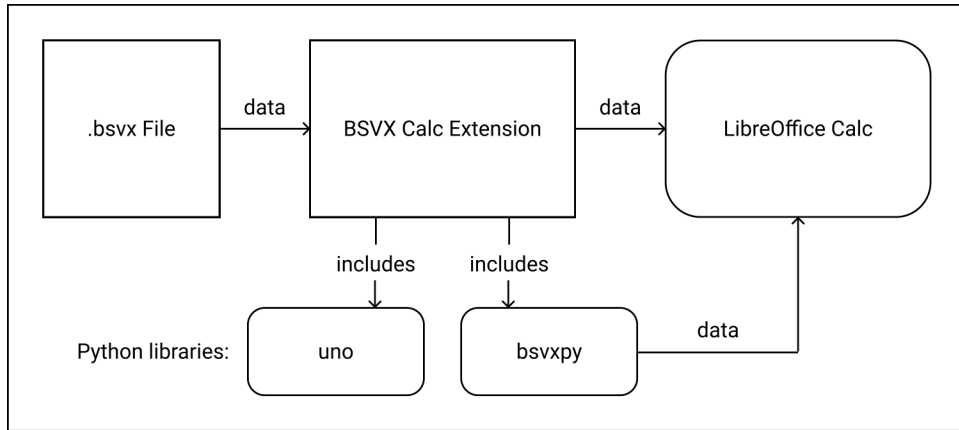

Figure 6: Scope of API and library calls for the BSVX Calc Extension in importing data from a `.bsvx` file.

This functionality also allows the user to convert from `.csv` to `.bsvx`. If the user chooses to import a `.csv` file into LibreOffice Calc, and exports that file to a `.bsvx` file, that same data will be available in both files. Likewise, if the user chooses to import a `.bsvx` file and export that file to a `.csv` file, that same data will be available. These changes do not significantly affect the overall architecture of LibreOffice Calc. The same base functionality of LibreOffice Calc is still provided to the user once this extension is installed, with the addition of importing and exporting features for `.bsvx` file formats.

Approach

As the specifications for the `.bsvx` file format have already been articulated, our main goal is to implement novel import/export functionality in Libre Calc. Example `.bsvx` files will be manually created to accurately account for testing basic functionality as well as edge cases. Initial builds will emphasize basic functionality such as correctly reading/writing basic data types, and will move to nested and blob types after initial testing is completed. Lastly, we'll attempt to optimize memory and computational efficiency to improve performance outcomes.

One important test of the `.bsvx` format and its corresponding extension will be its capacity to maintain full integrity of the data after conversion to and from `.csv` format. Our deliverable should enable a user with minimal experience in LibreOffice and some experience with spreadsheets to perform these conversions without loss of data or corruption of its ordering. Some loss of formatting and style is acceptable as long as the order of the values and the values themselves are maintained.

The project is expected to take approximately two months to complete, with production wrapping in late April 2020. Costs are to be kept minimal, if nonexistent, as our developers are being paid in "experience."

Active Developement

It should be noted that development of the `bsvxpy` library will require the use of a python virtual environment to ensure the user does not install the incomplete module on their default python environment. Specifics regarding the initialization and launch of a python virtual environment can be found in the README portion of the `bsvxpy` github.

Schedule and Milestones

Date	Goal
February 14, 2020	Finish initial background research, Draft BSV Proposal deliverable.
February 26, 2020	Finalize project's architecture, draft Project Architecture deliverable
March 4, 2020	**Midterm Milestone:** Present progress and evaluate stretch goals
March 18, 2020	Implement csv/bsvx backend
April 1, 2020	Draft the Initial Results deliverable
April 27, 2020	**Final Milestone:** Submit Final Report and Present findings to class

Table 4: Purported project schedule.

Challenges and Risks

One of the primary risks is the possibility that there are undiscovered ambiguities in the format specification. These will be dealt with by tightening the specification to account for ambiguities, and updating the reference implementation. Additionally, parsing `.csv` files for conversion to `.bsvx` files, and vice-versa, could involve numerous pitfalls. While there is an agreed upon standard format for `.csv` files, it didn't come about until 2005 and many `.csv` files still do not conform to it strictly. This will complicate our attempt to ensure integrity and continuity between conversions for *all* `.csv` and `.bsvx` files. For instance, when converting a a `.bsvx` file consisting of several strings of comma characters, our library would have to ensure that none of the commas end up being misinterpreted as delimiters. Properly following the specification will ensure consistency and prevent this from happening, but rigorous testing with a myriad of files will be necessary.

One challenge fundamental to the `.bsvx` format itself is deciding how to handle blob objects. It isn't always clear what type of data they should be deserialized as. Calc may contain methods to interpret unknown fields upon reading the file but more research on this is necessary. Another challenge is that the framework/API for both LibreOffice and Calc is unfamiliar, and it will take some time to learn.

Glossary

.bsvx The file extension associated with the Binary Separated Value file format.

.csv Comma-separated values file format often used for databases and spreadsheets.

.tsv Tab-separated values file format used for databases and spreadsheets.

Binary Separated Value The filetype associated with the encoding protocols laid out in this document.

Comma delimiter Practice of using the ',' character as a field separator to differentiate records in a file. Instances of a comma are always interpreted as a delimiter unless they appear in doubles quotes e.g. "1,0".

Deserialization The process of decoding of a .bsvx byte stream and reconstruction of the original data.

Libre Calc An open-source application for manipulating spreadsheets. Developed and maintained by The Document Foundation.

Serialization The process of encoding an object as a .bsvx format byte stream.

References

[1] Coleman, Larry. "Why do we keep using CSV?" Software Engineering Stack Exchange, 14 Feb. 2011. `https://softwareengineering.stackexchange.com/questions/47838/why-do-we-keep-using-csv`

[2] "CSV File Reading and Writing." Python Docs, 12 Feb. 2020. `https://docs.python.org/3/library/csv.html`

[3] Daly, James and Meiners, Chad. "Binary Separated Value." UMass Lowell, 30 Jan. 2020.

[4] Guthrie, Gordon. "How to Work With LibreOffice Calc." TechRadar, 23 July 2012. `https://www.techradar.com/news/world-of-tech/roundup/how-to-work-with-libreoffice-calc-1089870`