Principles of Data Design in Spreadsheets

Alexander Prutzkow

***Abstract*—Spreadsheets remain a relevant data processing tool for end users, despite the proliferation of databases and information systems. There are principles for writing easily-modifiable programs. However, there are no such principles for spreadsheets. We have formulated three principles for data design in spreadsheets. The data elementarity principle states that any component of a spreadsheet (cell, row or column, table, sheet) must contain indivisible (for this component and problem) data. This principle determines the arrangement of data in cells, tables, sheets, and spreadsheets. The data consistency principle states that data must not have contradictory values. This principle defines the relationship of data among themselves, the relationship of source and derived data. The principle, together with the previous principle, is related to the organization of data. The data certainty principle states that any component of a spreadsheet must have purpose. This principle determines the presentation of data in a workbook. The data must have names and a single designation. Each principle has rules that govern the details of data design in spreadsheets. Compliance with these principles and rules will make the spreadsheets readable and easily-modifiable. The formulated principles are used by us in spreadsheets for organizing the educational process and maintaining electronic journals, as well as in teaching how to work in spreadsheets.**

***Keywords*—Data design, data organization, principles, spreadsheets.**

I. What Inspired Us to Formulate Principles or Introduction

Any data processing can be programmed. A spreadsheet is a way to process data without programming (or with minimal programming). Therefore, spreadsheets are used by end users.

We teach spreadsheets at the School of Programmers of the Russian State Radio Engineering University (RSREU) [1] and Ryazan State Medical University. Students are required to explain how to arrange data on the sheet and why one organization of data is better than another.

We have already formulated the principles of easily-modifiable programs [2–3]. These principles allow you to say why one program is written better than another. Programs must follow code conventions as well.

All this inspired us to the purpose of the study.

II. Purpose of the Study

The purpose of the study is to formulate the principles for data design in spreadsheets so that spreadsheets are readable and easily-modifiable.

Manuscript received February 14, 2023.

A. Prutzkow is with the Ryazan State Radio Engineering University, 390005, Gagarin str., 59/1, Ryazan, Russia, with Ryazan State Medical University, 390026, Vysokovoltnaja str., 9, Ryazan, Russia, and with Lipetsk State Pedagogical University, 398020, Lenin str., 42, Lipetsk, Russia (e-mail: mail@prutzkow.com).

III. Related Works

A spreadsheet software developer has provided guidelines for organizing and formatting data in spreadsheets [4–5].

In [6], the rules for organizing data are formulated. These rules apply to individual cells as well as spreadsheets in general.

Practical guidelines (or rules) for organizing data in spreadsheets are discussed in [7–11]. In [12], the rules of using spreadsheets in institutions are proposed. In [7, 12], the rules are called the principles.

There is a software tool to automate the search for violations of the principles of using spreadsheets [13]. To check the types of source data and debug types, software engineering approaches can be used [14]. Templates of data arrangement in spreadsheets are highlighted in [15].

Errors in spreadsheets are explored and classified in [16–18]. Compliance with the rules of data organization in spreadsheets allows you to avoid unwanted data changes [19].

In [20–24], the best practices for arranging data in datasets are discussed. These practices can be used to organize data in spreadsheets. The rules of data organization should be taught to students when studying spreadsheets [25].

Why people use spreadsheets, what operations they perform and what problems they face is explored in [26–28].

These sources allowed us to formulate the following principles.

IV. Terminology

To formulate the principles, we introduce the following terminology.

A. Spreadsheets and Sheets

The term ‘spreadsheet’ means:

- ‘workbooks’ as a file with multiple ‘worksheets’ or;
- spreadsheet software.

The term ‘sheet’ means an individual ‘worksheet’.

B. Container and Components

A container is a collection of named components (see table).

The container has the following properties:

- boundedness: for each component it should be clear whether it is included in the container or not;
- naming: for each component within the container, it must be clear what kind of data it holds.

CONTAINERS AND THEIR COMPONENTS

CONTAINER	COMPONENT
Spreadsheet	Sheet
Sheet	Table or list as a table
Table	Rows (or columns)
List	Item (in row or column)
Row or column	Cells

C. *Entity and Object*

An object is a collection of parameter values.

An entity is a notion of a subject area that includes objects with the same parameters.

D. *Facts*

Facts are objects with the same parameters that have the following properties [29]:

- atomic: parameters are elementary;
- timestamped: one of the parameters is related to time;
- identifiable: the facts are pairwise distinct.

Facts form a list.

Fact is an object. However, let's single out the facts in a separate group and consider objects - all objects that are not facts.

V. Principles of Data Design in Spreadsheets

The principles of data design in spreadsheets are as follows:

- data elementarity principle;
- data consistency principle;
- data certainty principle.

Let's look at these principles and their corresponding rules in detail.

VI. Data Elementarity Principle

Data in spreadsheets must be elementary or atomic. Elementarity means the indivisibility of data.

The purpose of the principle is to reduce the time complexity of data processing by eliminating parse operations and their repetitions.

We mean the elementary character of data for each component (see table).

Elementarity of data can be defined in different ways for visual representation of data (human readability) or to simplify program processing (machine parsing).

The following rules apply to this principle.

A. *Data in cells must be elementary*

Elementary data does not need to be further separated. They can be combined in any order. For example, we can combine the last_name and first_name as

last_name + first_name

or

first name + last name.

B. *One table row – one entity object or one fact*

By extracting data from a table row, we know exactly how the object's parameters are located. To get a parameter, you only need a cell reference, not a formula to find and extract its value from a row or table.

Objects can also be arranged according to the columns of the table.

C. *One table – one entity or one list of facts*

When setting a table range in a formula, we must be sure that the table contains only objects of the same entity, but not several. This will simplify formulas for obtaining values.

Parameters of objects or facts in a table must be in the same order.

D. *One sheet – one table*

One sheet should contain only one table. This allows you to quickly find tables, copy tables along with a sheet to another spreadsheet.

E. *Do not use numbers in formulas*

The formula determines the order in which the value is calculated, and does not store the data. If you want to use a number in a formula, you must write the number in a cell and use a reference or a name in the formula.

VII. Data Consistency Principle

Every datum in spreadsheets must have the only one value.

The purpose of the principle is to eliminate the inconsistency of the data.

The following rules apply to this principle.

A. *All derived data must be computed*

If the derived data is not calculated but written as values, then when the source data changes, the derived data will need to be rewritten. If derived data is calculated, then when the source data changes, the derived data also changes.

B. *Data must not have duplicates in the same spreadsheet*

The main problem with duplication is a synchronization of the values of duplicates. The absence of duplicates removes this problem.

C. *If the input data has restrictions, then use the data validation function*

If the input data has restrictions, then it is necessary to check their values. Invalid values will skew the derived data. This will force you to check the values anyway, if not on input, then on calculation.

D. *If the number of cell values is limited and known in advance, then make a list of possible values to fill the cells*

«Preserve as much of the user's work as possible» [30]. The user can enter valid values, but with typos or errors. The drop-down list excludes such erroneous input (fig. 1).

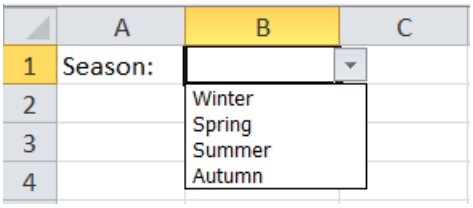

Fig. 1 – Select the season from the dropdown list

E. *Don't use formatting as derived data*

Suppose we need to detect abnormal values. We can highlight them with a background color or a column with an additional boolean type parameter (fig. 2). The TRUE value for this parameter indicates that the value is normal, and the

FALSE value indicates that the value is not normal.

	A	B
1	Weight	is Normal
2	75	TRUE
3	112	FALSE
4	54	TRUE

Fig. 2 – Discovering abnormal values by background color and additional parameter

The additional parameter allows you to filter and sort the values in contrast to the background color.

VIII. Data Certainty Principle

Data in spreadsheets must have a clear purpose.

The purpose of the principle is to improve the readability of data.

The following rules apply to this principle.

A. There must be no empty components in the container

An empty component is the same to the TODO comment in code: the data has not yet been received or it was forgotten to enter.

However, filling all components reduces the visibility of rare values. In this case, we recommend using conditional formatting.

B. All data on the sheet must be named

Can you tell us what data is shown in fig. 3?

	A	B	C
1			
2		142	
3			

Fig. 3 – Nameless number in cell

It could be:

- number without label or;
- the number that should have been removed or;
- part of the remaining label to another number.

The correct answer can be any of the above. Therefore, the data on the sheet must be named.

C. Follow to the naming conventions

The naming conventions contain the rules for naming containers and their components (see table). Look for such conventions in your institution before you start naming. If there are no such conventions, then briefly and clearly describe it on a separate sheet of the spreadsheet.

D. Data must have a single designation

At the beginning of the article, we introduced the terms. The terms allowed us to unambiguously state the text of the article, and you were allowed to unambiguously understand it. Data designations should be clear to both the creator of the spreadsheet and its users.

E. Use parentheses in formulas

The parentheses in the formula are used for the following:

- determine the order of operations;
- highlight the logical parts of the formula.

Deep nesting of parentheses reduces readability. Unfortunately, spreadsheets don't have a way to break down formulas like a program can be broken down into subroutines.

F. Entity parameter values must have the same format

A different data format can be in the following cases:

- a single format is not applied to the parameter values or;
- values are not an entity parameter due to incorrect data organization.

To remove this ambiguity, entity parameter values must have the same format.

G. Tables must have outer and inner borders

The border does not have to be a frame. The border can also be a different background color.

H. Use names for constants, not cell references

Spreadsheets allow you to access cells not only by reference, but also by name. Use names for constants to make the formula more readable.

IX. Discussions in Questions and Answers

Q: What influenced the formulated principles?

A: The formulated principles were influenced by:

- theory of relational databases;
- principles of object-oriented programming;
- code conventions;
- principles of graphical user interface design.

Q: Does it follow from rule A of the data certainty principle that there should not be empty cells on the sheet outside the table?

A: Yes.

Q: What are the criteria for data design in spreadsheets?

A: The criteria for data design in spreadsheets are readability and easily-modifiable.

Q: Are the principles formulated exhaustive?

A: At the current stage of the study, yes. But we do not leave out the emergence of rules that require the formulation of a new principle.

Q: What can a cell contain?

A: Based on the above rules, a cell can contain one of the following:

(1) elementary datum or;
(2) formula or;
(3) label or description.

X. Conclusions

We have formulated three principles for data design in spreadsheets:

(1) data elementarity principle;
(2) data consistency principle;
(3) data certainty principle.

The proposed and already known rules are manifestations of these principles. By following these principles and rules, the spreadsheet will be readable and easily-modifiable.

We use spreadsheets in the following cases:

- managing the operations of the School of Programmers of the RSREU for the calculation of indicators for various periods, the formation of reporting documents;
- electronic journals of students for counting missed classes, marks received points and the formation of final grades.

We use both spreadsheet software: standalone application and online services. Spreadsheets in online services are shared with other lecturers and dean.

In our practice, we adhere to the formulated principles. These principles will be used in the educational process of the School of Programmers of the RSREU and the Ryazan State Medical University.

REFERENCES

[1] A. Prutzkow, Napravlenija sovershenstvovanija dejatelnosti Gorodskoj shkoly programmistov RGRTU [Directions for Improving the Activities of the School of Programmers in RSREU]. In Aktualnye problemy sovremennoj nauki i proizvodstva, 2021:216-219.

[2] A. Prutzkow, Criterion and Principles of Easily-modified Program. In Workshop on Materials and Engineer. in Aeronautics, IOP Conf. Series: Materials Science and Engineer., 2021, 1027, 012025, DOI: 10.1088/1757-899X/1027/1/012025.

[3] A. Prutzkow, Tonkosti Programmirovanija v Primerah [Programming Subtleties in Examples]. Kurs, 2022.

[4] Guidelines for organizing and formatting data on a worksheet. URL: https://support.microsoft.com/en-us/office/guidelines-for-organizing-and-formatting-data-on-a-worksheet-90895cad-6c85-4e02-90d3-8798660166e3. Last accessed 2022/11/07.

[5] Top ten ways to clean your data - Microsoft Support. URL: https://support.microsoft.com/en-us/office/top-ten-ways-to-clean-your-data-2844b620-677c-47a7-ac3e-c2e157d1db19. Accessed 2023/01/03.

[6] K. Broman, K. Woo, Data Organization in Spreadsheets. In the American Statistician, 2018, 72(1):2–10, DOI: 10.1080/00031305.2017.1375989.

[7] A. Barbieri, Productivity Through Data Simplicity: Your Guide to Data Organization and Standardization in Excel. Amazon Digital Services LLC - KDP Print US, 2019.

[8] P. Bewig, How Do You Know Your Spreadsheet is Right? Principles, Techniques and Practice of Spreadsheet Style, 2005.

[9] J. Raffensperger, New Guidelines for Spreadsheets. In EuSpRIG, 2001.

[10] D. Clough, Excel Tips for Chemical Engineers | AIChE, 2017. URL:https://www.aiche.org/chenected/series/excel-tips-chemical-engineers. Accessed 2023/01/03.

[11] D. Limoges, Data Management. Excel Tips and Tricks to Summarize Data, Presentation, 2021. URL: https://www.ichpnet.org/events/annual_meeting/2021/ce/082_-_Limoges_Daniel_-_Data_Management_Excel_Tips_-_1up.pdf. Accessed 2023/01/03.

[12] M. Izza, Twenty Principles for Good Spreadsheet Practice, 3rd ed. ICAEW, 2018.

[13] Principles of good Excel use // Excel Advanced Training // PerfectXL. URL: https://www.perfectxl.com/online-excel-training/principles/. Accessed 2023/01/14.

[14] M. Erwig, Software Engineering for Spreadsheets. In IEEE Soft., 26:25–30, DOI: 10.1109/MS.2009.140.

[15] R. Teixeira, V. Amaral, On the Emergence of Patterns for Spreadsheets Data Arrangements. In Federation of Int. Conf. on Soft. Tech.: Applications and Foundations, Springer, 2016:333-345.

[16] E. Dobell, S. Herold, J. Buckley, Spreadsheet Error Types and Their Prevalence in a Healthcare Context. In J. of Organizat. and End User Comput., 2018, 30:20-42. DOI: 10.4018/JOEUC.2018040102.

[17] S. Powell, K. Baker, B. Lawson, Errors in Operational Spreadsheets. In J. of Organizat. and End User Comput., 2009, 21(3):24-36.

[18] K. Rajalingham, D. Chadwick, B. Knight, Classification of Spreadsheet Errors. In Proc. of the EuSpRIG Annual Conf., 2000:23–34.

[19] J. Strand, Error Tight: Exercises To Prevent Mistakes. In Psycholog. Methods, 2022, DOI: 10.1037/met0000547.

[20] D. Isbell, Open Science, Data Analysis, and Data Sharing. In L. Plonsky (ed.) Open Science in Applied Linguistics, 2021.

[21] H. Wickham, Tidy Data. In J. Statist. Soft., 2014, 59(1):1–23, DOI: 10.18637/jss.v059.i10.

[22] P. Soranno, Six Simple Steps to Share Your Data When Publishing Research Articles. In Limnology and Oceanography Bulletin, 28, DOI: 10.1002/lob.10303.

[23] K. Horstmann, R. Arslan, S. Greiff, Editorial Generating Codebooks to Ensure the Independent Use of Research Data: Some Guidelines. In Eur. J. of Psycholog. Assessment, 2020, 36(5):721–729, DOI: 10.1027/1015-5759/a000620.

[24] S. Ellis, J. Leek, How to Share Data for Collaboration. 2017, DOI: 10.7287/peerj.preprints.3139v5.

[25] F. Tort, Teaching Spreadsheets: Curriculum Design Principles. 2010.

[26] C. Chambers, C. Scaffidi, Struggling to Excel: A Field Study of Challenges Faced by Spreadsheet Users. In 2010 IEEE Symp. Visual Lang. and Human-Centric Comput.:187–194, 2010, DOI: 10.1109/VLHCC.2010.33

[27] L. Bartram, M. Correll, M. Tory, Untidy Data: The Unreasonable Effectiveness of Tables. In IEEE Transactions on Visualization and Computer Graphics, 2021, DOI: 10.1109/TVCG.2021.3114830.

[28] G. Chalhoub, A. Sarkar, “It’s Freedom to Put Things Where My Mind Wants”: Understanding and Improving the User Experience of Structuring Data in Spreadsheets. In CHI Conf. on Human Factors in Comput. Systems (CHI’22), ACM, 2022, DOI: 10.1145/3491102.3501833.

[29] N. Marz, J. Warren, Big Data. Principles and Best Practices of Scalable Real-Time Data Systems. Manning, 2015.

[30] W. Galitz, The Essential Guide to User Interface Design: An Introduction to GUI Design Principles and Techniques, 3rd ed. Wiley, 2007.