Intuitiveness as the Next Stage of Open Data: dataset design and complexity

Sarazin Arthur[ad*], Mourey Mathis[bd], Debru Romain[c]

[a]Researcher in Design Science, Datactivist, France

[b]Researcher in Data science, The Hague University of Applied Sciences, Netherlands

[c]Researcher in Social Marketing, Tour Graduate School of Management, France

[d]Researcher associated with the UNESCO Chair "AI and Data Science for Society"

[*]Corresponding author: Sarazin Arthur; arthur@datactivi.st

Abstract

Intuitive open datasets, which can adapt to the level of data literacy and the needs of the user, represent the next stage of open data. They not only provide broader access to data, but also unlock the underlying information, knowledge, and reuse potential. In this practice paper, we present a conceptual meta-design framework for designing new intuitive datasets and redesigning existing ones. This framework is the first output of the Dataflow research project, which aims to empower more data users to extract value from open data. Our framework is flexible and can be applied to any new or existing dataset to enhance its intuitiveness. Through this paper, we contribute to the open data community by offering a practical approach to designing and redesigning intuitive datasets and advancing the state of openness.

Keywords: open data; meta-design; data literacy; framework

Author roles:

Sarazin Arthur
Roles : Conceptualization, Methodology, Visualization, Writing - Original draft.

Mourey Mathis
Roles : Conceptualization, Formal analysis, Software

Debru Romain
Roles : Validation, Writing - Review editing

1 Overview

Repository location Recherche.data.gouv.fr

Context This conceptual meta-design framework was produced as part of the research project Dataflow, which aims to demonstrate how design can assist in producing useful open datasets and related data products. The framework and the associated prototype was created to support the open data community in their efforts to broaden the number of data publics (Ruppert, 2012) who use data to support decision-making, create new services and products, or produce innovative information and knowledge (Safarov, Meijer, & Grimmelikhuijsen, 2017).

The conceptual meta-design framework meets the need for a method to design intuitive datasets, that is datasets whose shape can adapt to the data literacy level and the need of the user. For their is a very diverse range of data users whose data literacy and needs differ greatly considering data : some will only look for one information in the dataset while other will use data as a core artifact of a data product they are making. Yet, these diverse needs and data literacy have not been considered by open data producers while designing datasets (Dymytrova, Larroche, & Paquienséguy, 2018).

The associated tool (currently in alpha prototype version) we propose will support the open data community in their effort to design new open data set or redesign existing ones whose reuse can address societal issues such as global warming, health and public transparency among others.

2 Method

To create the conceptual meta-design framework we used the design science research methodology (Hevner, March, Park, & Ram, 2004) and applied the Hierarchical design pattern (Vaishnavi & Kuechler, 2015, p.136). Such pattern uses the divide and conquer strategy to design a complex system. It design a system (the conceptual meta-design framework) by decomposing it into subsystems (five conceptual design framework to design each of the five level of abstraction of one dataset), designing each of them before designing the interactions between them.

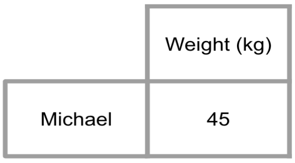
We also made sure, following the recurvise principle of granular computing that secures a high level of human-data interaction (Wilke & Portmann, 2016), that datasets of a upper level of abstraction could be constructed by human extrapolation of datasets of lower level of abstraction

We consider five level of abstraction for every dataset, hence five conceptual design design framework :

- **Level 4** of abstraction and design framework assist in designing data made of **unlinkable and multi-level datasets**
- **Level 3** of abstraction and design framework assist in designing data made of **linkable and multi-level datasets**
- **Level 2** of abstraction and design framework assist in designing data made of **a single dataset with several entities and attributes**
- **Level 1** of abstraction and design framework assist in designing data made of **a single entity and several attributes or of a single attribute and several entities**
- **Level 0** of abstraction and design framework assist in designing data made of a single entity, a single attribute and a single value. Level 0 corresponds to the classical definition of a data as a triplet entity-value-attribute (Redman, 1997) and is also considered as **the fundamental information granule**.

2.1 (Level 0) Designing data made of a single entity, attribute and value

Data made of a single entity, attribute and value corresponds in our framework to a level 0 dataset, with no complexity. It can also be referred to as a ”datum”. A ”datum” is defined as the smallest informational granule or the fundamental particle of data science. In computer science, the datum is defined as a triple entity-attribute-value. It can be represented by a table with a single cell. (see Figure 1).



	Weight (kg)
Michael	45

Figure 1: Table view of level 0 data

We have chosen to represent each entity in the following geometric shape (see Figure 2).

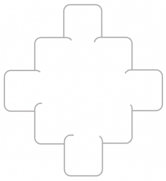


Figure 2: Graphical representation of an entity

This entity is defined by multiple attributes represented as subdivisions of this geometric shape (see Figure 3).

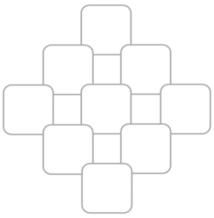


Figure 3: Graphical representation of an entity with several attributes and associated values

To each of these attributes corresponds a value, which we will name w, x, y, and z (see Figure 4)

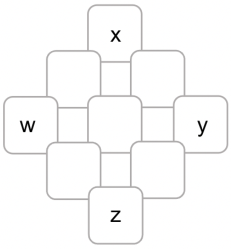


Figure 4: Graphical representation of an entity with several attributes and associated values

Using these graphical conventions, we represent the level 0 data as follows (see Figure 5)

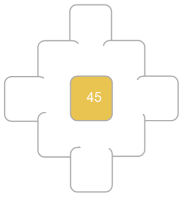


Figure 5: Graphical representation of a level 0 data or datum

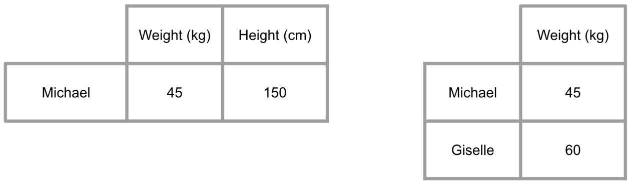
At the lowest abstraction level, level 0, data has no complexity as no interpretation is required for its comprehension. This type of data is commonly referred to as "raw data." For instance, in our example, Michael weighs 45 kg, which is a fact.

At the level of abstraction 0, data is the prerogative of machines, which cannot interpret it automatically but store the datum in their memory, awaiting its use.

2.2 (Level 1) Designing data made of a single entity and several attributes or with a single attribute and several entities

.

To move from level 0 to level 1 of abstraction, we need to find a common element to several datum. It can be a common attribute (‘weight‘), a common entity (‘michael‘) or a common value. As a result we obtain a table with a single entity defined by multiple attributes, or a table with a single attribute and multiple entities (see Table view below in Figure 6). This table is the primary form of what we call "data" (plural of "datum")



	Weight (kg)	Height (cm)
Michael	45	150

	Weight (kg)
Michael	45
Giselle	60

Figure 6: Table view of level 1 data

Level 1 data can be represented graphically as follows (see Figure 7)

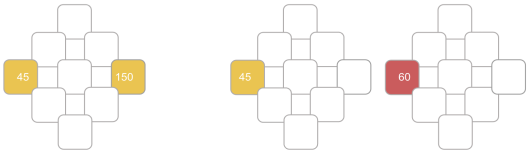
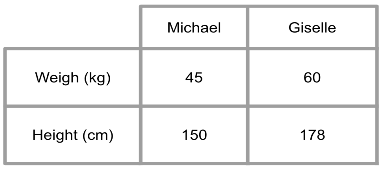


Figure 7: Graphical representation of level 1 data

2.3 (Level 2) Designing data made of a single dataset with several entities and attributes

.

To increase complexity and move from level 1 to level 2, it logically involves adding attributes and/or entities to the table (see Figure 8)



	Michael	Giselle
Weigh (kg)	45	60
Height (cm)	150	178

Figure 8: Table view of a level 2 data

We can graphically represent this table as follows (see Figure 9)

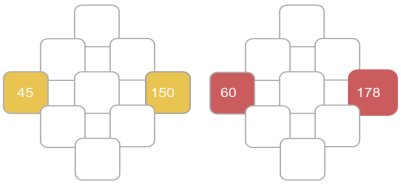
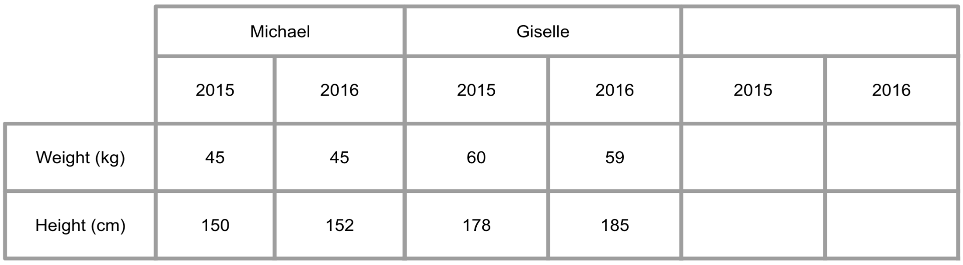


Figure 9: Graphical representation of a level 2 data

2.4 (Level 3) Designing data made of linkable and multi-level datasets

.

At a higher level of abstraction, data transition from one to many linkable datasets and from one to many levels of entities and/or attributes. It results in linkable datasets, consisting of multiple levels of entities and multiple levels of attributes (multi-level linkable datasets see Figure 10).



	Michael		Giselle			
	2015	2016	2015	2016	2015	2016
Weight (kg)	45	45	60	59		
Height (cm)	150	152	178	185		

Figure 10: Table view of a Level 3 data

In the above example, we have two levels of entities: individuals on one side, and years on the other. We also have one level of attribute: individual physical characteristics (weight and height).

We also provide a graphical representation of this level of abstraction (see Figure 11)

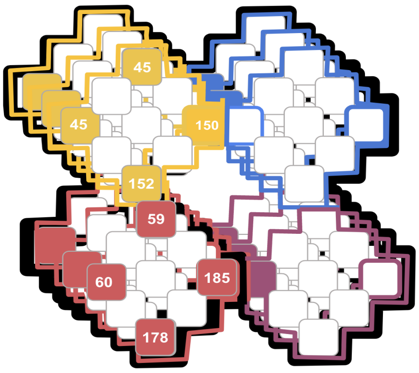
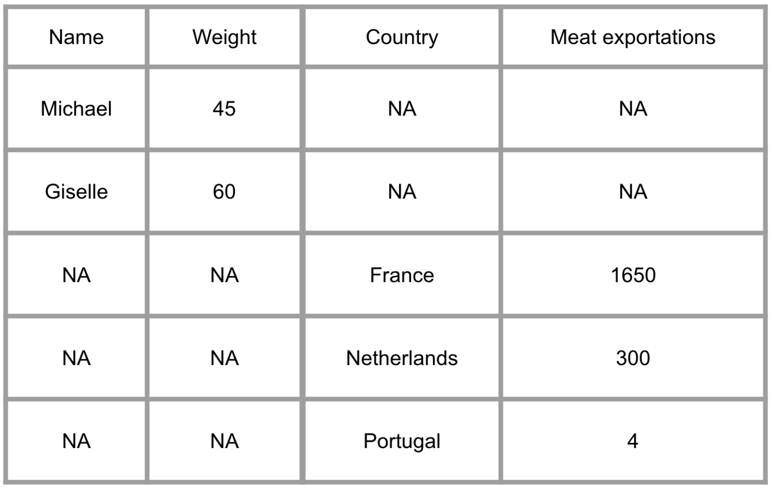


Figure 11: Graphical representation of a level 3 data

2.5 (Level 4) Designing data made of unlinkable and multi-level datasets

.

At a higher level of abstraction, data transition from definable complexity to undefinable complexity. That is multi-level data tables that cannot be linked based on current scientific knowledge. These multi-level tables are characterized by the fact that their junction cannot be represented in the form of a table, since their level of complexity is indefinable. At best, they could be represented by an amalgamation of two tables whose connections are not available (see Figure 12 below).



Name	Weight	Country	Meat exportations
Michael	45	NA	NA
Giselle	60	NA	NA
NA	NA	France	1650
NA	NA	Netherlands	300
NA	NA	Portugal	4

Figure 12: Table view of a level 4 data

This very high level of abstraction still remains the prerogative of human beings, who are capable of connecting data with links whose complexity is indefinable, thanks to concepts. It can be represented graphically as follows (see Figure 13)

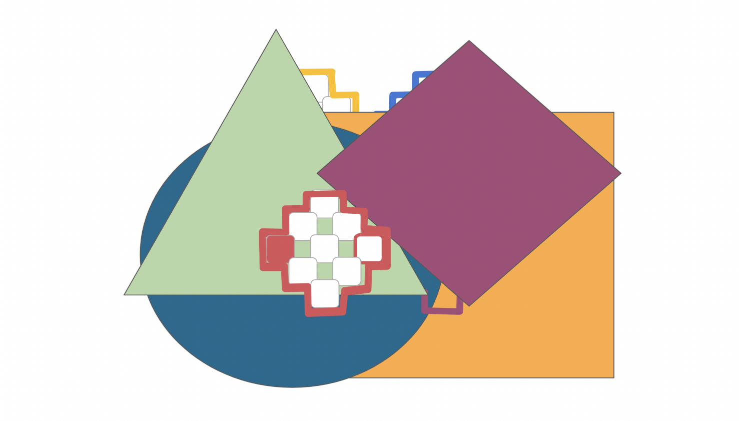


Figure 13: Graphical view of a level 4 data

It has become commonplace to state that there is an ever-increasing amount of data available, which implies an underlying structure of extreme intelligence that can link data together. This structure would enable the discovery of fascinating knowledge and the creation of innovative applications.

In this article, we assume that the reality of newly available data is quite different: there is no apparent structure to link them together, or if it does exist, it is indefinable based on current scientific knowledge. **From our conceptual meta-design framework's perspective, all newly available datasets are level 4 data.**

Indeed, the available data, in their vast majority, do not share any attributes or common elements that would allow them to be linked together. In this regard, the "data gold rush" of newly available data can be compared to a set of circles, rectangles, triangles, but also words, numbers, letters that, because they are placed on the same visual plane, would have an underlying structure that we could learn a lot from (see Figure 13). This is possible. However, in most cases, these links are of indefinable complexity based on current scientific knowledge. For example, the link between daily shopping baskets in supermarkets and monthly water consumption in surrounding households has not been established. It may not exist, but certainly the complexity of the link is indefinable based on current scientific knowledge.

We have shown in this section that data can be represented according to 5 levels of abstraction. Level 0 and level 4 are purely theoretical in nature, being respectively the domain of machines and human beings. Apart from constructing a complete and relevant classification, we will not be interested in these theoretical levels in the rest of our article. Our main focus is to establish the conceptual framework of intuitive data, which, as a reminder, are defined as data whose level of abstraction and complexity can truly (and not theoretically) adapt to the data literacy level of its user.

In the following section, we will formally demonstrate that transitioning from a higher abstraction level to a lower level does indeed decrease the complexity of the data. We will also determine to what extent this complexity decreases.
With this formal demonstration, we prove that our conceptual meta-design framework can be used to design datasets that can adapt to the level of data literacy and needs of the user. **That is that it can be used to design intuitive open datasets.**

3 Data set complexity

Let's start with a formal definition of dataset complexity :

Definition of dataset complexity:
The complexity of a dataset can be measured by the number of relationships that can be extracted from it. In this framework, we consider that the order of complexity ($\mathcal{C}()$) associated with a dataset relates to how fast the complexity increases with the size of the dataset.

Let us consider the derivation of the order of complexity for each level of complexity below:

Level 0:
The complexity of a single data point has to be equal to zero. There is no complexity associated with a single point of information. The order of complexity is $\mathcal{C}(0)$.

Level 1:
For a variable or a single vector of values, there is only one way to interpret the data: we look at how all data points compare to each other (you could think of a line plot for time series or a barplot for cross-sectional data). Since there is only one way to look at the data, the order of complexity is $\mathcal{C}(1)$.

Level 2:
For a (single-level) table, we can consider the following:

(1) We can interpret each row/column independently.
(2) We can combine one row (or column) with one or more rows (or columns) to study the relationship they have (the information created from running a regression on multiple variables could be an example here).

Then we know that the overall number of combinations we can make in a table with n rows is: $2^n - 1$. We can see here that the complexity grows with the number of rows. Hence, we define the order of complexity as how fast the complexity grows with each new row:

$$2^{n+1} - 1 - (2^n - 1) = 2^{n+1} - 2^n = 2^n(2 - 1) = 2^n$$

The order of complexity is then: $\mathcal{C}(2^n)$

Level 3:
For a multi-level table, the complexity depends on the number of rows/columns (n) but also the

number of groups for each level (g). We have then a number of total possible combinations: $2^{ng} - 1$. If we then consider the growth of complexity every time we add a new group to a level:

$$2^{n(g+1)} - 1 - (2^{ng} - 1) = 2^{n(g+1)} - 2^{ng}$$

$$2^{ng+n} - 2^{ng} = 2^{ng}(2^n - 1)$$

We find the order of complexity to be: $\mathcal{C}(2^{ng}(2^n - 1))$

Level 4: Few unlikable tables = complexity 4 ($\mathcal{C}(\infty)$)

4 Complexity reduction

In this section, we take a look at how to reduce the order of complexity of a dataset by transforming a higher level of complexity into a smaller one. We only consider jump of 1 level of complexity down. Specifically, we look at the relative reduction in order of complexity: $\Delta_{\mathcal{C}} = \frac{\mathcal{C}_{after} - \mathcal{C}_{before}}{\mathcal{C}_{before}}$

4.1 Complexity reduction from level 4 to level 3

.

Going from an infinitely complex dataset to a measurably complex dataset is, by definition, an **almost perfect reduction of complexity**.

4.2 Complexity reduction from level 3 to level 2

.

For the upper bound of the reduction, we consider the most complex dataset of the level consider, hence, we look at the limits of the reduction when the size of the dataset shoots up to infinity.

$$\Delta_{\mathcal{C}} = \lim_{g \to \infty} \frac{2^n - (2^{ng}(2^n - 1)}{(2^{ng}(2^n - 1)}$$

$$\lim_{g \to \infty} \frac{2^n}{(2^{ng}(2^n - 1)} - 1$$

$$\lim_{g \to \infty} \frac{1}{(2^g(2^n - 1)} - 1 = 0 - 1 = -100\%$$

Regarding the lower reduction bound, we need to consider the simplest level 3 dataset. The simplest is a multi-level table with 2 groups (g) and 2 attributes (or entities) (n) only. For such a dataset, the reduction to a level 2 gives:

$$\Delta_{\mathcal{C}} = \frac{1}{2^g(2^n - 1)} - 1 = \frac{1}{4(4 - 1)} - 1 = -91.\overline{6}\%$$

With our definition, **the reduction of a level 3 dataset to a level 2 is bounded such that:** $\Delta_{\mathcal{C}} \in [91.\overline{6}\%; 100\%[$

4.3 Complexity reduction from level 2 to level 1

.

$$\Delta_{\mathcal{C}} = \lim_{n \to \infty} \frac{1 - 2^n}{2^n}$$

$$\lim_{n \to \infty} \frac{1}{2^n} - 1 = 0 - 1 = -100\%$$

Similarly as above, for the lower reduction bound, we consider the simplest level 2 dataset. The simplest is a table with only 2 attributes (or entities) (n). For such a dataset, the reduction to a level 1 gives:

$$\Delta_{\mathcal{C}} = \frac{1}{2^n} - 1 = \frac{1}{4} - 1 = -75\%$$

Hence, the reduction of a level 2 dataset to a level 1 is bounded such that: $\Delta_{\mathcal{C}} \in [75\%; 100\%[$

4.4 Complexity reduction from level 1 to level 0

.

$$\Delta_{\mathcal{C}} = \frac{0-1}{1} = -100\%$$

The computation is trivial for the transition from level 1 to level 0. **Any level of complexity reduced to 0 corresponds to a 100% reduction in complexity.**

5 Method and Dataset Instantiation

Object name Data redesign method

Format names and versions Pseudocode

Creation dates 2023-09-01

Dataset creators Mathis Mourey (programmer, data editor), Arthur Sarazin (designer, data designer)

Language Gherkin language (Python style)

License MIT License.

Repository name https://github.com/datactivist/data_redesign_method

Publication date 2023-08-30.

We decided to use pseudocode and the Gherkin language so the data redesign method can be both human and machine-readable, and can, as a result, be instantiated both in technical artefacts (such as a python library) and in design patterns.

6 Reuse Potential

This pseudocode can be used by designers, data scientist and enlightened citizens who deal with real world data. More importantly, the fact that the pseudocode can be translated into a formal set of sequence (using Python for example) implies that international open data platforms such as UIS.stat or the World Bank Open Data can use it to design and code a *data redesign plugin* that will increase the intuitiveness of their datasets.

This would give birth to surprising *raw data reuses*, that are reuses that can still be grasped by many data public and yet being more human-based than raw data.

Acknowledgements

The authors would like to express their deepest gratitude to Datactivist and the UNESCO Chair "AI and Data Science for Society" who supported and encouraged the development of this redesign method and invested so that public institutions can translate it into social implact.

Funding Statement

This research resulted from the research project Dataflow funded by Datactivist and the UNESCO Chair in AI and Data science for Society

Competing interests

The author(s) has/have no competing interests to declare.

References

Dymytrova, V., Larroche, V., & Paquienséguy, F. (2018). *Cadres d'usage des données par des développeurs, des data scientists et des data journalistes livrable n°3* [Research report]. Retrieved from https://hal.science/hal-01730820 (Citation Key: dymytrova:hal-01730820 tex.hal$_{i}d : hal - 01730820 tex.hal_{v}ersion : v1$)

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *Management Information Systems Quarterly*, *28*(1), 6.

Redman, T. C. (1997). *Data quality for the information age.* Artech House, Inc. (Citation Key: redman1997data)

Ruppert, E. (2012). Doing the transparent state: Open government data as performance indicators. In *A world of indicators: The production of knowledge and justice in an interconnected world* (p. 51–78). Cambridge University Press.

Safarov, I., Meijer, A., & Grimmelikhuijsen, S. (2017). Utilization of open government data: A systematic literature review of types, conditions, effects and users. *Information Polity*, *22*(1), 1–24.

Vaishnavi, V. K., & Kuechler, W. (2015). *Design science research methods and patterns: innovating information and communication technology.* Crc Press.

Wilke, G., & Portmann, E. (2016). Granular computing as a basis of human–data interaction: a cognitive cities use case. *Granular Computing*, *1*, 181–197.