Data organization issues in civil engineering

Marian Smutný, Jiří Kaiser
Czech Technical University in Prague – Faculty of Civil Engineering, Thákurova 7, 166 29 Praha 6, Czech Republic
marian.smutny@fsv.cvut.cz

Abstract. In today’s organization of data in civil engineering there remains ambiguity, diversity of terminology and consequent lack of clarity. The daily reality is to work with hybrid data sources in a variety of software programs. In CAD programs it is possible to import and convert elements from GIS, but they will be classified only into layers. Or, attach external references from CAD files in BIM, but there still remains the problem of finding your way around the confusing list of layers, and their abbreviations. Also broadly used filesystems as place for store and organise data on disk into files (entities) and hierarchical structure of folders (categories) is limited by reductionist categorization approach. Another problem shows with absence of multilingual approach. The limits are in the methods of data organization, and so only partially portable. Therefore, if there is a common denominator to all problems, it is the absence of a unifying organizational environment. In current computer applications used in civil engineering, methods of data classification are implemented that do not allow multiplicative or faceted categorization and thus inevitably create an antagonistic environment with any previous or future classification system. Unfortunately, even if a software tool implements a modern multiplicative categorization or domain ontology, it will not significantly solve the overall situation, because it will still not be possible to use such a method of categorization in other applications. Therefore, if there should exist a solution that would significantly facilitate cooperation and orientation in the data, it must inevitably work independently, across diverse software applications, with the ability to link the functions of the application with such a separate categorization system.

1. Introduction
After 10 years of applied research in design practice and analysis of critical points, the issue of data sorting has proved the most critical. There are various purposes of classification. E.g. in order to apply level of detail principles [1] or level of development principles [2] [3] for entities of BIM we need to classify them. Also classification of objects in BIM libraries is necessary [4].

Subsequent reorganization of the categorization, change of names, and possible typing errors leads to data entropy. For this reason, there is a reluctance to change, update and correct file names, directory structure, and layer organization, etc. Likewise, user groups regroup and rename the original structure irreversibly complicating the orientation of the original group of users and disrupting the interconnection of data.

More than 20% of working time of knowledge workers is devoted to information retrieval, while 62% are dissatisfied with the search result [5]. According to my observations, the reorganization of data when taking over documents, restructuring and renaming, when submitting to the contracting authority,
takes up to 40% of working time for designers. This is why it is important to deal with this particular issue.

Each user group usually prepares its structure to organize data, such as names and categorized entities [6].

Depending on the type of data it is:

1) Graphical elements [7]
   1) CAD elements constituting the schematic construction drawings, user-categorized into a flat layer list.
   2) BIM elements representing spatial building structures, categorized in a reductionist hierarchical structure according to the host application.
   3) Attributes attached to graphical elements, usually consisting of the attribute's own name and value.

2) Documents
   1) Usually stored as files categorized in the reductionist directory structure of the file system (abbr. FS [8]). Content management systems (CMS) and document management systems (DMS) are used in particular only inside corporations
   2) Filename is considered as a unique identification document.

These structures differ partially or completely between groups. If these groups want to cooperate, they have three options:

1) conversion - means import/export, i.e. reorganizing the data at every transfer.
   o laborious for all parties involved

2) adaptation to heterogeneous data organization in every single project where they co-operate.
   o difficult to train, remember and follow.

3) unification - i.e. enforcement of a uniform standard of data organization.
   o problematic to enforce a standard that suits the majority.
   o too large and therefore confusing having to cover the needs of different jobs and different processors.
   o inflexible - to customize is disallowed, to meet emerging needs.

Before processing the data, it is necessary to navigate inside received documents, take over the relevant data, rename and move them according to the current standard of the processor. This reorganization cancels the previous organisation.

1.1. Categorization systems
To resolve this situation, various corporations and interest groups are proposing unification, that is, formal enumerative systematization. Although more advanced methods of data categorization have been known for 100 years (Ranganathan), they routinely choose a 400-year-old reductionist enumerative taxonomy [9] (Linné) based on systematic names or codes (e.g. SNIM [10]). Working with such a structure presupposes knowledge of these codes and their error-free writing.

Formal enumerative systematization construction activities, products and materials (such as exists in a chemical or biological taxonomy), as well as attempts at a uniform structure and coding elements and activities in the construction industry, is rapidly becoming obsolete. "Because it is an enumerative system (that is, the main goal is to find a place for each subject, rather than to build a coherent structure)" [11] The reason for this obsolescence is that in the natural sciences is systematized status-quo, while human ingenuity is dynamic. "Enumerative systems are really easy to work with and are the best choice for a static universe of knowledge. Unfortunately, there is no universe of knowledge that is static." [12]
It is excluded in the reductionist categorization:
1) Multiplicative assignment of an entity to two or more categories.
   1) According to the various aspects (doors are the object of wall openings, but may or may not be a facade element)
   2) Controversial categorization of an entity (e.g. a French window or a balcony door belong to the categories of windows and doors at the same time)
2) Multilingual labels of the same entity, category or attribute - by different names depending on the language, industry practices and corporate methodology.

2. Basic principles
If we have uncategorized entities, it is time-consuming to locate and select them. The task is much simpler if entities are categorized. Categories organized into a tree hierarchical structure are most natural for human understanding, organization of knowledge and visualization of information [13]. Commonly used principles of currently used classification systems are subject of various technical standards [14] (e.g. [15]).

When real-world entities are categorized, we rely only on the reductionist hierarchy. Locations (cabinet, shelf, briefcase, box, etc.) represent categories, and screws represent entities. At the outset, it is necessary to think how the individual aspects will be arranged hierarchically (e.g. Type>Head>Drive>Material or Material>Type>Head>Drive etc.). The entity is then categorized into only one eligible category. Later, should you need to select all the steel entities, you should need to go through the whole structure in depth, and mark all the steel categories. A possible reorganization of the structure would cost too much effort.

Unfortunately, in the design of today's information (also classification), systems remain an anachronism, restricting their behavior by limiting the material world. However, the virtual world of information systems, unlike the material world, allows one to search for entities much faster, keep entities in virtual connections, classify an entity according to various aspects, and also allow one to maintain multiple names for an entity or category (synonyms, multilingualism, abbreviations).

While reductionist categorization allows only one entity to be in one place at a time, multiplicative categorizations allow the so-called colocalization of entities in one space-time. The multiplicative categorization is possible that in one space-time has been several different entities. This brings the concept of faceted classification. "A faceted approach to the organization of knowledge has been discussed among experts throughout the 20th century. In practice, however, the facet principle seemed to be problematic for a long time - it was difficult to use for the linear physical arrangement of e.g. books on a shelf and for the visualization of such a system of knowledge organization. However, in the current virtual network environment of the website, technological conditions are created for the use of its advantages."[16]

3. Document management
The goal of data organization is to quickly find and select relevant data. A hierarchical model is currently used for the data organization system, usually applied in the file system (abbreviation FS) or in the database of DMS.

Before the rise of computers, the term file system (FS) referred to the procedures and methods for handling paper documents. Since 1958, the term has been implemented in computers in a similar sense. Categories (i.e. directories, folders), group entities (i.e. files), and subcategories (i.e. subdirectories, subfolders) on disk, and are used to allow the user to organize them hierarchically. For addressability reasons, file system has 3 basic limitations (except for FS that supports hard link):
   1) There can be only one path to an entity.
2) An entity cannot be in more than one category at a time.
3) There cannot be two entities or subcategories with the same name in one category.

When working on various architectural projects, there may be a need to change, expand, refine or regroup the structure to include new facts. Implementing changes and reworking the obsolete structure is often laborious, and after restructuring, complicates pathfinding for existing users. For this reason, the restructuring of the categorization is unpopular, so those relics remain in the structure, the ambiguity increases, and over time will prevent effective location of the data.

3.1. Limitations of reductionist categorization
The current approach will be illustrated by the example of two documents: "2.FP CW1.1 electric r1.0" and "1.FP CW2.1 water r1.0" categorized in the directory structure. Directory/folder names represent categories into which documents (entities) are categorized. The directory structure forces the user to prioritize only one aspect of classification, so in the second layer of categorization there is redundant use of the same named categories (either ARCHITECTURE - MEP&HVAC - STRUCTURES or CW01 - CW02). There are many options, here are two options: localization preference leading to the redundancy of specialization (Table 1), or Specialization preference leading to the redundancy of localization (Table 2).

### Table 1. Localization preference leading to the redundancy of specialization.

| localization preference | redundancy of specialization |
|-------------------------|-----------------------------|
| [PROJECT 0123]          | [PROJECT 0123]              |
| [CW1]                   | [CW1]                       |
| [ARCHITECTURE]          | [ARCHITECTURE]              |
| [MEP&HVAC]              | [MEP&HVAC]                  |
| "2.FP CW1.1 electric r1.0" | "2.FP CW1.1 electric r1.0" |
| [STRUCTURES]            | [STRUCTURES]                |
| [CW2]                   | [CW2]                       |
| [ARCHITECTURE]          | [ARCHITECTURE]              |
| [MEP&HVAC]              | [MEP&HVAC]                  |
| "1.FP CW2.1 water r1.0" | "1.FP CW2.1 water r1.0"     |
| [STRUCTURES]            | [STRUCTURES]                |

### Table 2. Specialization preference leading to the redundancy of localization.

| specialization preference | redundancy of localization |
|--------------------------|---------------------------|
| [PROJECT 0123]           | [PROJECT 0123]            |
| [ARCHITECTURE]           | [ARCHITECTURE]            |
| [CW1]                    | [CW1]                     |
| [CW2]                    | [CW2]                     |
| [MEP&HVAC]               | [MEP&HVAC]                |
| [CW1]                    | [CW1]                     |
| "2.FP CW1.1 electric r1.0" | "2.FP CW1.1 electric r1.0" |
| [CW2]                    | [CW2]                     |
| "1.FP CW2.1 water r1.0"  | "1.FP CW2.1 water r1.0"   |
| [STRUCTURES]             | [STRUCTURES]              |
| [CW1]                    | [CW1]                     |
| [CW2]                    | [CW2]                     |
Such redundancy in the next name-change often leads to the omission of the change, in all their occurrences. If the name is changed, it is necessary to search for, and replace all occurrences of the name in bulk. However, in the case of the division of the structure, by introducing more detailed subcategories, e.g. by dividing the MEP&HVAC category into the ELECTRIC and WATER subcategories, such an adjustment cannot be made by a simple mass exchange. Inconsistent adjustments then lead to entropy of the whole structure.

External references are most sensitive to preserving the name and structure of the directory. The external reference is immersing whole or part of the source document, into one or more documents. The result is a complete document, integrally displaying content from other documents. If the content of the external source changes, these changes are immediately reflected in all occurrences. External references refer to the file name along with a relative or absolute directory path. Changing the file name, directory name, or moving it to another directory will cause the link to be lost (Figure 1).

![Figure 1. External References Manager. The yellow icon indicates files for which the link has been lost.](image)

Therefore, it is not possible to change the file name or directory structure without losing the link, even in cases where the typed errors need to be corrected. Such a change would require a change to be recorded in different programs in different data formats. In addition, there is no central record of where, and at how many places, the file is referenced.

It is obvious that the same meaning can be given by another group of users by a different name (e.g. Building Services Engineering vs. Building Equipment, or Structural Analysis vs. Structural Engineering) (Figure 2).
Different user groups need different names for the same entity (Table 3):

1) **systematic name** *SysName* - that most accurately identifies category
2) **abbreviation** *Abbr*
3) **alphanumeric code** *Code*
4) **trivial name** *NickName*

| *UsrGrp*   | *lng* | *SysName*                  | *Abbr*   | *Code*   | *NickName* |
|------------|-------|----------------------------|----------|----------|------------|
| CTP        | en    | Building Equipment         | MEP&HVAC |          | Systems    |
| CS-KPS     | en    | Technical Installations    | MEP&HVAC | I        | Installations |
| SKANSKA    | en    | Building Services Engineering | BSE        |          | Services   |

4. **Content Management**

The current CAD programs are able to classify a graphic element only in a single layer. Layers are organized in either a flat, simple list (e.g. *AutoCAD*) or a reductionist hierarchy (e.g. *MicroStation*, *Allplan 100*). This precludes classifying elements according to inter-twinning aspects.

CAD elements are graphic entities (lines, circles, hatches, ...) that schematically represent the building structure. Ideally, each such element would be categorized according to the building structure it represents. For example, all elements that together represent dividing walls (thin and thick lines, dashed and solid, hatching, etc.) should be categorized in the category of *Vertical Nonbearing Interior* structures. At the same time, it should be possible to classify one and the same element into several categories (layers), e.g. a dashed line representing the edge of the door sill could be categorized in both the *Wall* layer and the *Door* layer.

4.1. **Limitation of simple lists**

An example is the elements showing the technical installation of buildings. These elements can be categorized according to two different aspects into 9 categories:

1) Aspect of **specialization**
   - LTG - electric light circuits
   - SO - electric socket circuits
   - ELV - extra low-voltage circuits
   - GAS - gas installations
- **WAT** - water installations
- **SEWS** - sanitary sewage

2) Aspect of schema
- **LIN** - lines (cable routing, pipes, etc.)
- **OBJ** - objects (switches, valve fittings, distributors, etc.)
- **DSC** - labels (text, annotation lines, arrows, dimensions, symbols, etc.)

For these needs, there is a multiple increase in layers in a simple list (Table 4):

| Table 4. Resulted list of required layers. |
|------------------------------------------|
| lines | objects | labels |
|------|---------|--------|
| **electric light circuits**               | LTG LIN | LTG OBJ | LTG DSC |
| **electric socket circuits**              | SO LIN  | SO OBJ  | SO DSC  |
| **extra low-voltage circuits**            | ELV LIN | ELV OBJ | ELV DSC |
| **water installations**                   | WAT LIN | WAT OBJ | WAT DSC |
| **sanitary sewage**                       | SEWS LIN| SEWS OBJ| SEWS DSC|
| **gas installations**                     | GAS LIN | GAS OBJ | GAS DSC |

If required to isolate only electrical elements from the drawing, it is necessary to mark item by item and switch off layers of all other installations (WAT LIN, WAT OBJ, WAT DSC, SEWS LIN, SEWS OBJ, SEWS DSC, GAS LIN, GAS OBJ, GAS DSC). Or for coordinating the installation, mark off all layers containing descriptions (LTG DSC, SO DSC, ELV DSC, WAT DSC, SEWS DSC, GAS DSC).

It is important to realize that in common practice, there is a need to use many more categories than stated here, and the number of layers thus increases quadratically.

The use of layers for categorization (in the sense of a simple list) has the following drawbacks:
1) It is not possible to simply select a group of similar layers. One must select one after another.
2) One cannot assign a single element to multiple layers.
3) The enumerative list of layer names (or codes) become too extensive and complex.
4) An error in the name (or code) of the layer, or by renaming the layer, will create problems during batch processing (e.g. for import or export)
5) Names must be unique (whilst context works in hierarchies)

4.2. Limitation of the reductionist hierarchy
Although some CAD programs allow layers to be grouped or organized hierarchically, their fundamental drawback is that layers are considered entities, and not categories in themselves. Even so, the same limitations mentioned above occur. The following example shows two ways (Table 5) (Table 6) in which layers containing **SO OBJ** (e.g. electric socket) and **WAT OBJ** (e.g. water meter) installation elements can be categorized into a hierarchical structure.
Table 5. Resource preference leading to the redundancy of schema.

| Resource preference | Redundancy of Schema                          |
|---------------------|----------------------------------------------|
| [electricity]       | [electricity]                                |
| [electric socket circuits] | [electric socket circuits]                      |
| [lines]             | [lines]                                      |
| [objects]           | [objects]                                    |
| SO_OBJ              | SO_OBJ                                       |
| [descriptions]      | [descriptions]                               |
| [extra low-voltage circuits] | [extra low-voltage circuits]                   |
| [lines]             | [lines]                                      |
| [objects]           | [objects]                                    |
| [descriptions]      | [descriptions]                               |
| [sanitary]          | [sanitary]                                   |
| [sanitary sewage]   | [sanitary sewage]                            |
| [lines]             | [lines]                                      |
| [objects]           | [objects]                                    |
| [water installations] | [water installations]                          |
| [lines]             | [lines]                                      |
| [objects]           | [objects]                                    |
| WAT_OBJ             | WAT_OBJ                                      |
| [descriptions]      | [descriptions]                               |

Table 6. Schema preference leading to the redundancy of resource.

| Schema preference | Redundancy of Resource |
|-------------------|------------------------|
| [MEP&HVAC]        | [MEP&HVAC]             |
| [lines]           | [lines]                |
| [electric socket circuits] | [electric socket circuits]                      |
| [extra low-voltage circuits] | [extra low-voltage circuits]                   |
| [sanitary sewage] | [sanitary sewage]     |
| [water installations] | [water installations]                          |
| [objects]         | [objects]             |
| SO_OBJ            | SO_OBJ                |
| [descriptions]    | [descriptions]        |
| [extra low-voltage circuits] | [extra low-voltage circuits]                   |
| [sanitary sewage] | [sanitary sewage]     |
| [water installations] | [water installations]                          |
| WAT_OBJ           | WAT_OBJ               |
| [descriptions]    | [descriptions]        |

CAD elements have been mentioned so far, but similar limitations exist with BIM elements. Although they are categorized in a hierarchical structure, this structure is reductionist. This persists as to whether, for example, a balcony door can be categorized into windows or doors. In the case of multiplicative categorization, they can be categorized into either both categories or their parent category. In similar cases, IFC [17] allows the use of general types of elements (eg. IfcBuildingElementProxy, IfcDistributionElement), which can practically be considered as a categorization into a parent category. One or more additional categorizations can be assigned to BIM elements using the
IfcRelAssociatesClassification categorization entity, but this cannot be considered a multiplicative categorization in the true sense of the word.

5. Results and discussions
In current computer applications used in civil engineering, methods of data classification are implemented that do not allow multiplicative or faceted categorization and thus inevitably create an antagonistic environment with any previous or future classification system. Reductionist’s hierarchy that is applied in most commonly used classification systems usually requires preference of one particular view and leads to data redundancy. Unfortunately, even if a software tool implements a modern multiplicative categorization or domain ontology [18], it will not significantly solve the overall situation, because it will still not be possible to use such a method of categorization in other applications.

6. Conclusions
If there should exist a solution that would significantly facilitate cooperation and orientation in the data, it must inevitably work independently, across diverse software applications, with the ability to link the functions of the application with such a separate categorization system. The question for future research is to set requirements on a new classification system to achieve more effective ways of work with categorized elements.

Acknowledgment(s)
This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS20/103/OHK1/2T/11

References
[1] V. Nývlt, and K. Prušková, “Building Information Management as a Tool for Managing Knowledge throughout whole Building Life Cycle,” IOP Conf. Ser.: Mater. Sci. Eng., vol. 245, 2017.
[2] K. Prušková, “Reducing failures rate within the project documentation using Building Information Modelling, especially Level of Development,” MATEC Web of Conferences, vol. 146, 2018.
[3] K. Prušková, and V. Nývlt, “Issue of Building Information Modelling Implementation into the Czech Republic’s Legislation using the Level of Development,” IOP Conf. Ser.: Mater. Sci. Eng., vol. 245, 2017.
[4] M. Dědič, “3D scanning and analysis of acquired data of historically and culturally significant objects referring to the work of Adalbert Stifter,” MATEC Web of Conferences, vol. 279, 2019.
[5] Delphi Group, “Information Intelligence: Content classification and the enterprise taxonomy practice,” Boston, 2004.
[6] K. Antoš, J. Čejka, V. Dobiáš, S. Jirá, T. Náhlik, and H. Pavličíková, “Studijní materiály pro obor informatika,” VŠTE, p. 93, 2019 (in Czech).
[7] J. Hlavenka et al., “Nový výkladový slovník výpočetní techniky,” Computer Press, ISBN 80-85896-13-3, 1995 (in Czech).
[8] G. A. Barnard, and L. Fein, “Organization and Retrieval of Records Generated in a Large-Scale Engineering Project,” Eastern Joint Computer Conference, str.59-63, 1958.
[9] L. M. Chan, “Library of Congress subject headings: principles and application,” Westport, Conn., ISBN 1591581540, 2005.
[10] Czech BIM Council, “Standard Negrafických Informací 3D Modelu”, [Online] 2021 [Accessed 21-04-09], available from: https://snim.cz/bim/about-snim.
[11] B. H. Kwasnik, and V. Rubin, “Stretching conceptual structures in classifications across languages and cultures,” Cataloging & categorization Quarterly, vol. 37(1/2), pp.33-47, 2003.
[12] M. P. Satija, and D. Martínez-Avila, “Features, functions and components of a library
categorization system in the LIS tradition for the e-environment,” *Journal of information science theory and practice*, vol. 3, p.68, 2015.

[13] M. Lima, “The Book of Trees: Visualizing Branches of Knowledge,” 2014.

[14] K. Pruskova, “Beginning of Real Wide us of BIM Technology in Czech Republic,” IOP Conf. Ser.: Mater. Sci. Eng., vol. 471, 2020.

[15] ISO 12006-2:2015, “Building construction — Organization of information about construction works — Part 2: Framework for classification,” International Organization for Standardization, [Online] 2015, available from: https://www.iso.org/standard/61753.html.

[16] H. Kučerová, “Ontologie organizace znalosti,” Charles University, Faculty of Arts, p.82, 2018.

[17] ISO 16739-1:2018, “Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries,” International Organization for Standardization, [Online] 2018, available from: https://www.iso.org/standard/70303.html.

[18] V. Nývlt, “Conceptual Meta Model for Building Information Modeling,” *J. Phys.: Conf. Ser.*, vol. 1680, 2020.