Standardized Data Integration in the AEC Domains – What does it take to succeed?

G Paskaleva[1], A Mazak-Huemer[2], S Sint[3], T Bednar[3]

[1] Research Unit of Business Informatics, TU Wien, Vienna, Austria
[2] Department of Business Informatics, JKU, Linz, Austria
[3] Research Unit of Building Physics, TU Wien, Vienna, Austria

E-mail: galina.paskaleva@tuwien.ac.at

Abstract. Data models are the backbone of digital information exchange, since they contain the data to be exchanged. Just as information requirements vary, so do data models – in level of detail, level of abstraction, and in domain coverage. These variations are the reason for the ease of communication between some data models and for the difficulty in communication between others. Regarding the Building Information Model (BIM) initiative, the IFC standard's data model has varying detail and abstraction levels and large domain coverage. In contrast, the Austrian ÖGG guideline, for example, has a consistent detail and abstraction level and focuses on a single domain – subsurface and tunnel modelling. In order for such data models to participate in loss- and distortion-free information exchange, a reliable translation via, e.g., third data models is necessary.

In this paper, we present formal criteria for distinguishing between semantics-carrying data models, such as IFC and ÖGG, and translating data models that provide reliable communication bridges between them, such as XML, CAEX and SIMULTAN. We will show that translating data models are an indispensable part of the data model infrastructure even within a single domain. In addition, we will derive the minimal set of attributes of such models and demonstrate their necessity on a use case from the subsurface engineering domain.

1. Introduction

Data exchange between and within domains has been indispensable to the Architecture Engineering and Construction (AEC) industries. The rapid development of Computer Aided Design (CAD) tools and standards over the last decades is a testament to that [1]. One of the most visible open standard for data exchange is the Industry Foundation Classes (IFC) [2]. The latest version, IFC 4.3.x[1] has integrated some of the semantics relevant to the infrastructure domain, including tunneling. Since the level of digitalization in the tunneling domain is currently developing rapidly [3], we will take this as an example for the challenges we face in data exchange. Figure 1 shows a very small excerpt of the IFC 4.3.x standard concerning geotechnics. The types with pattern background illustrate a differentiation between different strata in the subsoil, depending on the dominating material – solid as a diagonal pattern, gases as a dotted pattern, and water as a wave pattern. Those appear to be reasonable placeholders for further developments in the domain.

[1] https://technical.buildingsmart.org/standards/ifc/ifc-schema-specifications/

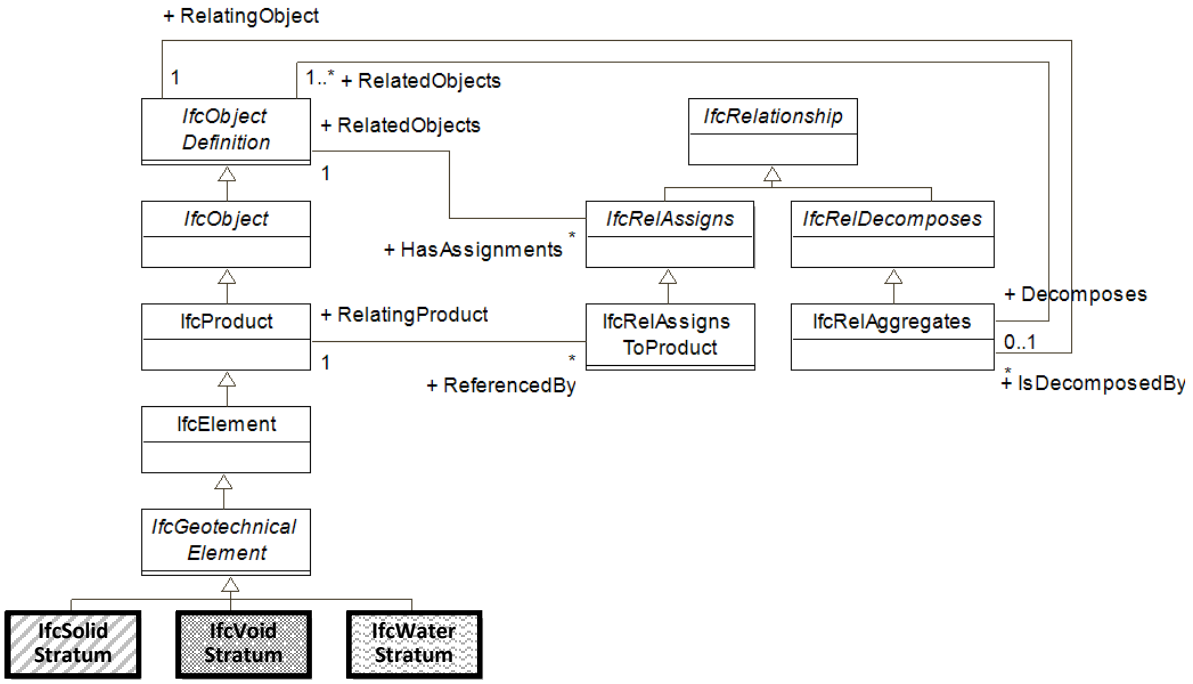

Figure 1. Excerpt of IFC 4.3.x demonstrating the differentiation in the geology, hydrogeology and geotechnics domains between solid, void and water strata

Let us take a look at another data model concerning the same subject, this time highly specialized and designed by domain experts in Austria, the ÖGG guideline [4].

Figure 2 shows an UML[2] class diagram we constructed from the guideline, which is written in a natural language. There are several significant deviations from the semantics contained in IFC (cf. Figure 1). There is no assumption that the subsoil consists of strata. Its structure is expressed in terms of behavior, geotechnical attributes and non-isotropic properties. The background patterns in Figure 2 show a tentative semantic correspondence to the IFC data model in Figure 1.

The remainder of the paper is as follows. In Section 2, we give a brief overview of different data models that are used in the AEC domain. In Section 3, we describe our approach and its application to data integration based on the two data models presented in this section. In Section 4, we present an overview of some related work. Finally, Section 5 concludes the paper and we give a brief outlook.

2. Background

In this section, we give an overview of extant data models both from a semantic as well as pragmatic, i.e., stemming from different interpretations of the same concept within the same domain [5], perspective. We examine their purpose and possible interplay in a data exchange scenario.

ÖGG and SIA. Both the ÖGG [4] and the SIA 199 [6] guidelines describe the subsurface domain with particular emphasis on geotechnics, geology and hydrogeology. The descriptions are in a natural language, which allows a lot of leeway in designing a corresponding data model,

[2] `https://www.omg.org/spec/UML/2.5.1/About-UML/`

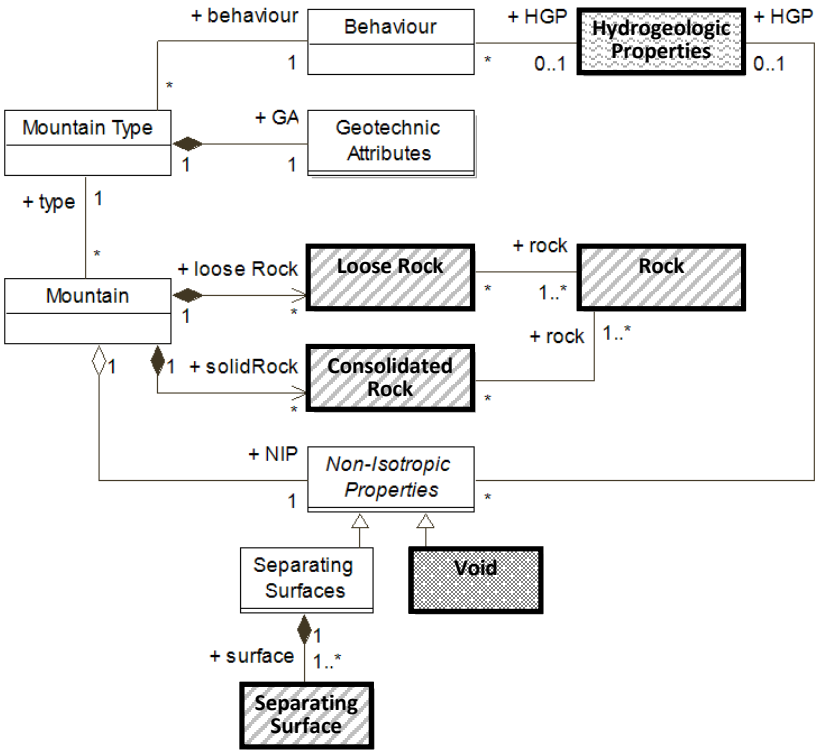

Figure 2. Excerpt from the Austrian ÖGG guideline concerning the geology, geotechnics and hydrogeology of the subsoil

as shown in Figure 3 and Figure 4 in the use case of discontinuity representation in ÖGG. This highlights one significant feature of semantic-carrying domain data models – they can be formalized in multiple complementary ways using different modeling languages.

IFC. The Industry Foundation Classes (IFC) [2] standard is an example of a data model that spans multiple domains, including architecture, structural engineering, HVAC, infrastructure, etc. Just as the previously described guidelines, it is concerned with semantics. However, in addition to that, it prescribes certain structure and naming convention dependencies. Those significantly limit the options when modelling a particular domain while, at the same time, allowing some creative freedom to the domain expert.

BIM. Building Information Modelling (BIM) is the overarching concepts that accommodates both the very specialized data models, such as ÖGG and SIA 199, as well as IFC and some even more generic ones. The idea behind BIM is a loss- and distortion-free data exchange spanning the entire life cycle of a built structure [7]. A possible end goal of BIM would be a fully operational digital twin that can both supply information to the physical structure and receive real-time feedback from it in all relevant domains [8]. Such complex infrastructure requires modularization and a strict separation of concerns in order to provide traceability and maintainability. A strict separation of the usage of data models, e.g., into semantic, communication, or translation data models, is one possible way of addressing these potential issues.

UML. The Universal Modelling Language (UML) offers a large library of diagrams for modelling various aspects of data, from semantics to data exchange and user interaction. It is

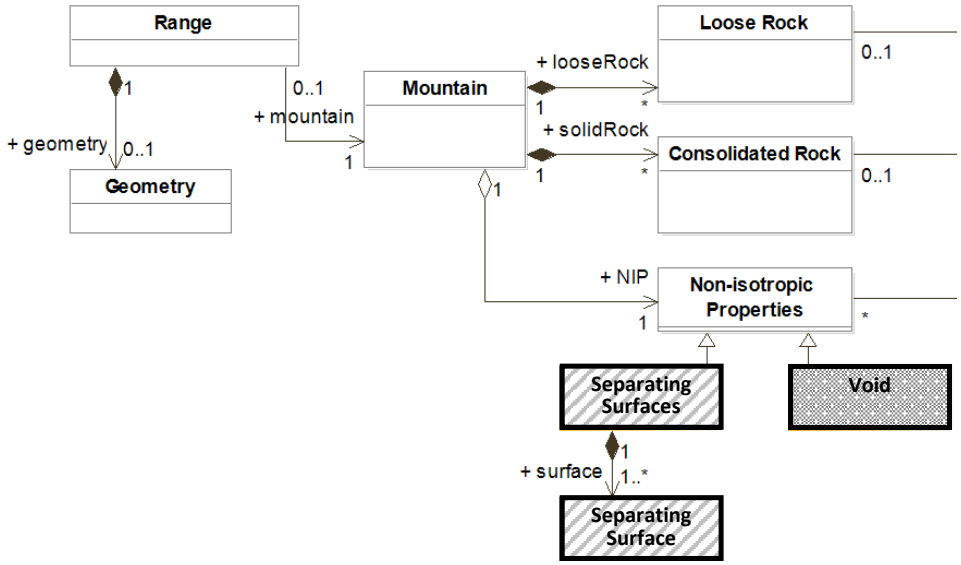

Figure 3. Excerpt of the ÖGG guideline as a UML model concerning discontinuities at a low semantic resolution

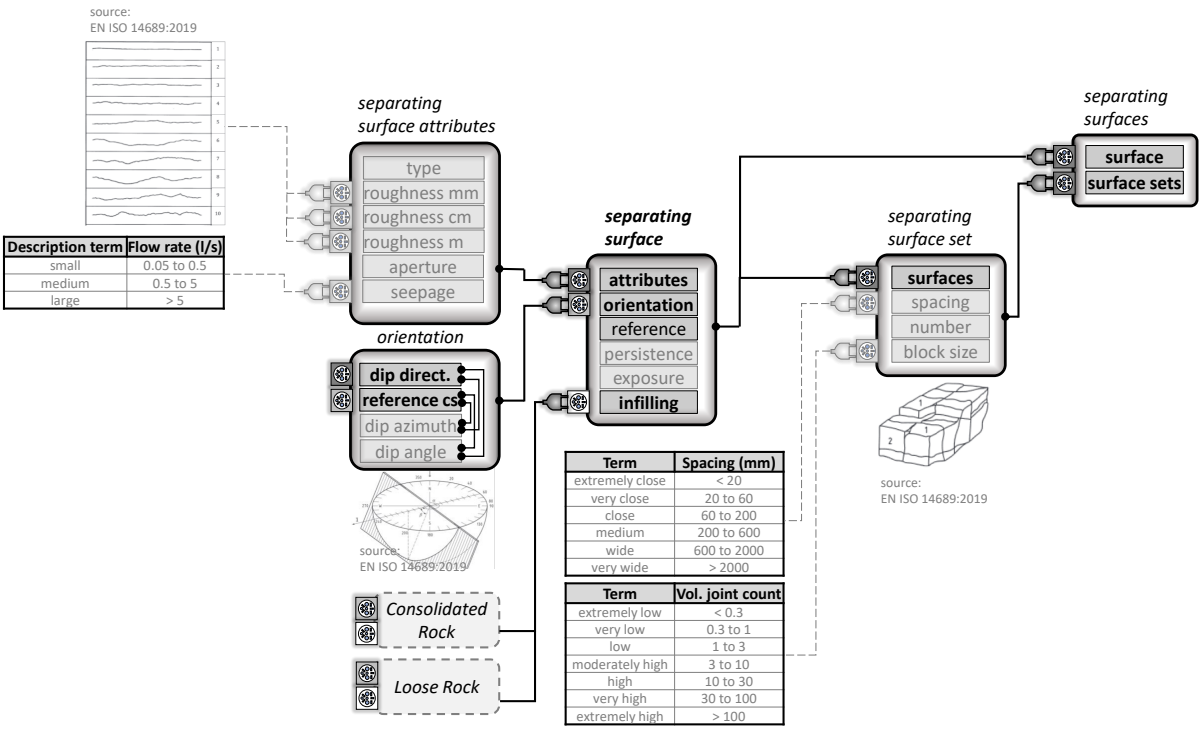

Figure 4. Excerpt of the ÖGG guideline as a CAEX model concerning discontinuities at a high semantic resolution

applicable to any domain that organizes its data in an object-oriented way, i.e., in well-defined chunks.

CAEX. Automation ML (AML[3]) is a standard for data exchange in automation system development [9]. It contains modules covering the architecture and general requirements (IEC62714-15), the class libraries for modeling engineering roles (IEC62714-26), and geometry and kinematics (IEC62714-37). As part of AML, the Computer Aided Engineering Exchange (CAEX) standard enables the modelling of physical and logical components for encapsulating different aspects in an engineering domain [9]. CAEX data objects enable the reuse of existing components with their roles and interfaces through cloning. Additionally, the hierarchical structure of CAEX allows the definition of arbitrarily complex models without loss of readability. CAEX is an example of a data model entirely dedicated to information exchange without imposing any semantic requirements. It is well suited to modelling the different aspects of an object as well as of communication flows.

SIMULTAN. This is an example of a generic object-oriented data model [10]. Similarly, to UML and CAEX, it can model both semantics and the data exchange between different semantic data models.

3. Research Methodology

Based on the overview given on the various data models used in the AEC domain, we now tackle the challenge of constructing a reliable communication bridge (standardized and efficient data integration) between semantic data models. For this purpose, we will use the example we introduced in Section 1, using the IFC and ÖGG data models, and concentrate on integrating information about strata in the subsoil.

When domain experts exchange information they employ taxonomies, physical and logical hierarchy and relationships with the primary goal of conveying knowledge, or semantics. Therefore, the models best suited to these requirements are the semantic data models. For example, a geotechnics expert is interested not just in structuring a mountain as a composition of units, but as a multi-faceted entity with different behaviors under different conditions. It is immaterial to the expert if this domain-specific knowledge is packed in a hierarchical model using composition or in a flat list of elements.

However, two domain experts can have diverging perspectives on the same domain, for example, in the context of an international tunneling project. In this case, their communication method and the underlying data model should not introduce additional semantics, just facilitate conversation, understanding and data integration.

In Section 1, we show the different semantic models concerning the subsurface domain in the IFC standard (cf. Figure 1) and the Austrian ÖGG guideline (cf. Figure 2), respectively. How do we integrate knowledge, some of which is modeled according to IFC and some – according to ÖGG? Do we produce a hybrid semantic model that integrates both perspectives on the same semantics? If another expert joins the project with knowledge modelled in another standard, e.g., the Swiss SIA 199 guideline, do we produce yet another hybrid semantic model?

Therefore, in order to facilitate both standardized and efficient data integration, i.e. the communication bridge, we need to answer the following research questions:

RQ1: If we integrate information from n semantic models, do we need a new hybrid semantic model?

[3] `https://www.automationml.org/`

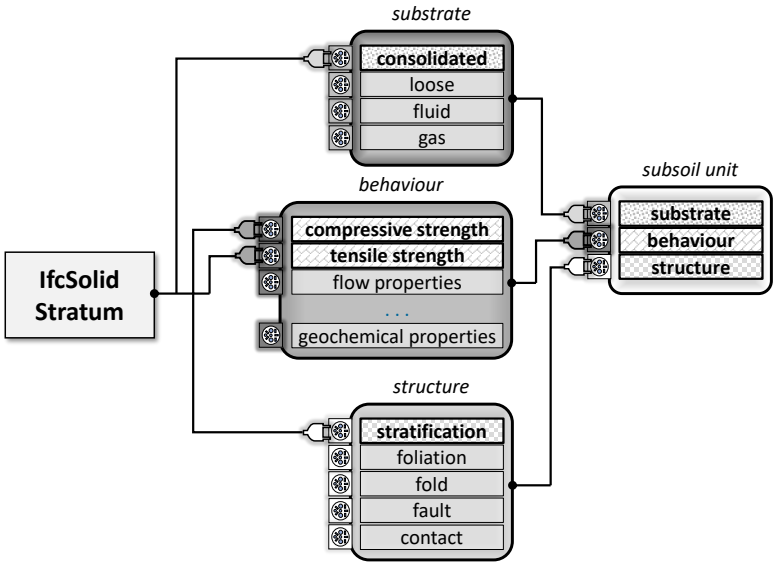

Figure 5. Reconstruction of the element *IfcSolidStratum* from the IFC 4.3.x standard relative to the common reference frame *subsoil unit*

RQ2: How do we adapt the communication every time any of the n semantic models changes or an additional one has to be included?

RQ3: How do we recognize contradictions and only partial semantic overlap?

A method that is capable of answering all of the above questions is reconstruction and communication [11]. It is a concept that enables communication between experts in the context of a common reference frame. In the following subsections, we will demonstrate its application on the motivating example (cf. Section 1) and will provide answers to the three research questions above.

3.1. Reconstruction

The step that enables mutual understanding between different perspectives on the same semantics is reconstruction. Figure 5 shows how this method can be applied to the Element *IfcSolidStratum* from the IFC model from Figure 1. First, the common reference frame (CRF) is defined, comparable to a common spatial coordinate system. In this case, the CRF is subsoil unit, which maps semantics along three independent axes – substrate, behavior and structure. The CRF has to be agreed on by all involved experts to make sure that it allows a complete mapping of each semantic model.

In the next step, we reconstruct the semantic element of interest as discrete positions along each axis of the CRF. In this case, the substrate of *IfcSolidStratum* can only be consolidated, its behavior can be described by compressive and tensile strength (among many others), and its structure is a stratification by its very definition.

We perform the same reconstructive step for as many elements of the other semantic model as necessary. The result is presented in Figure 6. As the ÖGG guideline is much more detailed, its semantics has finer resolution and we need to include multiple elements to correspond to every relevant aspect of *IfcSolidStratum* – *Consolidated Rock, Loose Rock, Behaviour*, and *Separating Surface*. The element *Hydrogeologic Properties* is included, since even solid strata have those. They were possibly omitted from this version of the IFC standard because it models

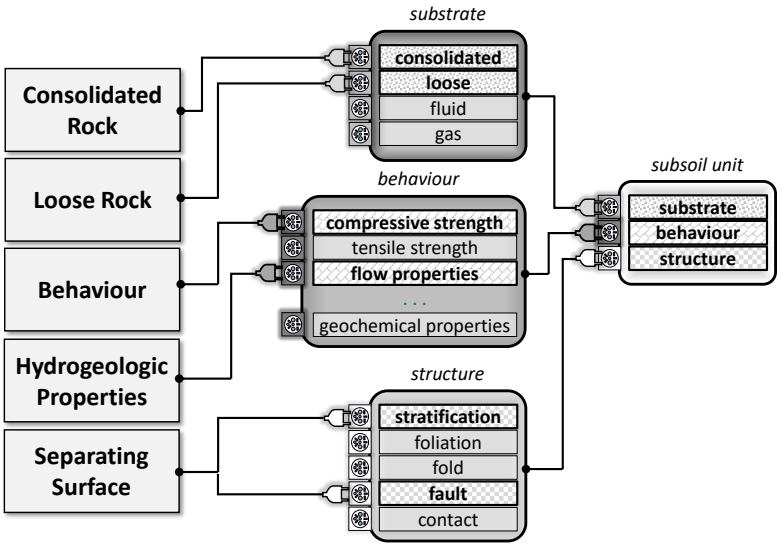

Figure 6. Reconstruction of the elements from the Austrian ÖGG guideline corresponding to *IfcSolidStratum* relative to the same common reference frame as shown in Figure 5

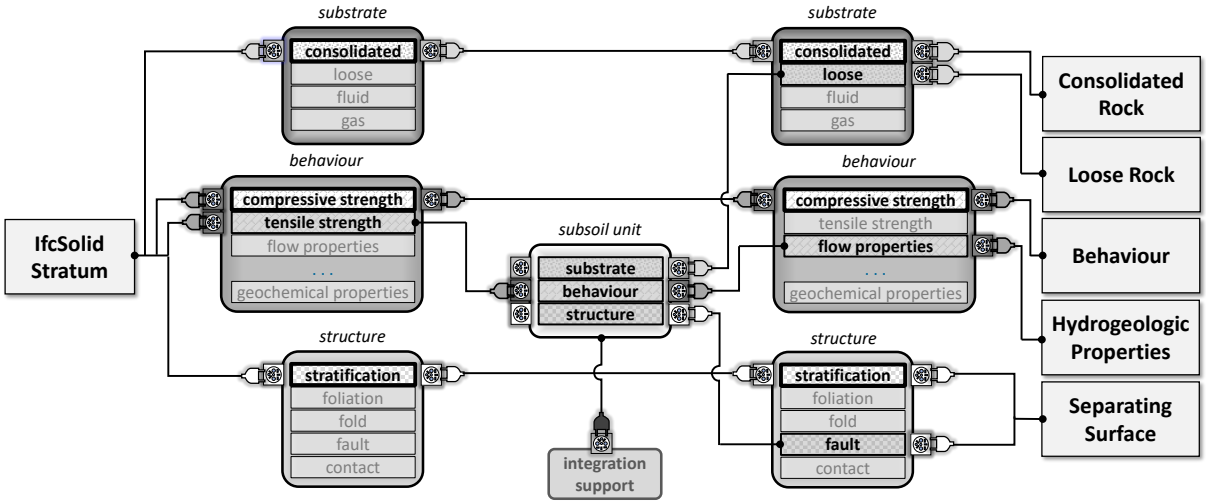

Figure 7. Integration of *IfcSolidStratum* from IFC 4.3.x with the elements from the more detailed Austrian ÖGG guideline within the same reference frame

the subsurface domain at a much lower resolution.

3.2. Communication

After the reconstruction of the relevant semantic elements is complete, it becomes possible to communicate the differences of perspective in the established CRF. This step is depicted in Figure 7. It becomes apparent that there is some direct correspondence, e.g., along the substrate axis we can translate the properties of the consolidated substrate directly.

However, there is also incomplete overlap, e.g., along the same axis, the properties of a loose substrate have no counterpart in *IfcSolidStratum*. This makes a loss-free translation impossible. Therefore, we have to perform integration instead, i.e., we assemble a fuller semantics from the two semantic models we have. In Figure 7, the boxes with thick black border represent the

directly translatable parts, the boxes with thin black border represent parts that need to be handled by an integration support unit (see bottom of the figure) so as not be lost or distorted during information exchange.

Based on these insights, we can now answer the RQs:

Answer RQ1: No matter how may semantic models we integrate information from, we need no hybrid semantic model, just a Common Reference Frame (CRF) suitable for all of them.

Answer RQ2: Any changes in the already included semantic models or the addition of new models can be handled by the reconstruction step in the same CRF. The only exception occurs when the newly added model requires the addition of a new axis to the CRF, which makes a repeat of the reconstruction step for all models necessary.

Answer RQ3: As we have already shown, contradictions and partial overlap between the semantic models produced by different experts in the same domain, are not uncommon. However, the reconstruction step makes both contradictions (e.g., strict separation between solid and fluid units is required or not) as well as partial overlaps (e.g., in the behavior) apparent, and gives the domain experts the opportunity to discover and handle them explicitly.

3.3. Data Model Classification

Most BIM applications offer export and import functionality in a commonly used data format, such as XML[4], CSV[5] or JSON[6], because these formats have underlying data models that act as generic containers for any semantics. It is of note that IFC does not fall under this category, since it incorporates semantics. For example, IFC knows what a wall is; XML does not, in spite of being able to transport its information.

One question remains: What model did we employ for the reconstruction and communication between IFC and the ÖGG guideline? Was it not also a semantic one, since we separated three semantically labelled axes, substrate, behavior, and structure? Let us consider the minimal requirements on this model's features. In essence, we need to be able to define an arbitrary number of axes with arbitrary names and mapping points on them. Further, we need to be able to connect these points in any way necessary. Examples for such data models include, but are not limited to, XML, CAEX, and SIMULTAN. What all those models have in common is that they are domain-independent and are specifically developed to be able to hold and connect semantics without adding anything of their own.

In summary, we need to distinguish between semantics-carrying and communication-enabling data models. The first can be of arbitrary complexity and structure, as required by the domain they model. The second should be of the lowest possible complexity, allowing only containers, labelling, and interconnectivity between elements. This is crucial in order to avoid semantic cross-contamination and to provide maximal clarity during the data integration process. Dictionaries and grammar textbooks do not invent additional languages in order to provide robust translation between existing ones. Similarly, semantic integration methods should avoid adding any semantics to the domains they serve.

4. Related Work

Several lines of research address data integration of specific tools. However, there is little work that examines the challenge of integration at a generic level, which is discussed below.

[4] `https://www.w3.org/TR/xmlschema11-1/`

[5] `https://datatracker.ietf.org/doc/html/rfc4180`

[6] `https://www.ecma-international.org/publications-and-standards/standards/ecma-404/`

Mayerhofer et al. [12] examine the offered data exchange of AutomationML, which on the one hand uses the existing industrial data format PLCopen XML and on the other hand offers an Intermediate Modeling Layer (IML) with appropriate transformation rules to decouple the used modeling languages from the target format PLCopen XML. Since IML and the transformation rules are only semi-formally described, the authors clarify syntactic and semantic aspects of IML and introduce a metamodel and operational semantics for IML. In contrast to our work, they propose the use of IML as a semantic domain for heterogeneous plant behavior models, thus processing semantic information in this layer rather than just creating a link. Paskaleva et al. [13] present another approach that also deals with data exchange using AutomationML. The authors address the challenges of implementing interoperability and seamless data exchange in a multidisciplinary collaboration between different stakeholders in civil engineering projects. They present a synthetic reconstructive approach to dealing with heterogeneous perspectives on the subsoil based on the neutral data format CAEX, while facilitating the integration of different data sources without limiting their autonomy or diversity. While in their approach the granularity of the data models is very similar, we consider different granularities. The approach of Stark et al. [14] goes in a different direction. The authors investigate a concept for cloud-based integration and exchange of data between engineering tools. In this context, they define interfaces for a repository on the one hand and AML-based importers and exporters for the RobotStudio tool on the other. The approach follows a realization for specific data structures, whereas we aim for a generic solution of interconnection.

In the AEC domain, another approach addresses the integration of geographic information systems (GIS) and building information models (BIM), which are mostly managed separately [15]. The authors propose a framework for data integration and simplification to improve site planning and building design, and propose an integrated BIM-GIS model with a multi-level data structure. In their approach, various parsing programs for common BIM and GIS data formats are developed to extract information to enable integration on a file and search graph basis.

Maass and Lampe [16] address data integration in a different area. Their approach is specialized and concerned with the provision of product data and describes an extended data model that integrates standardized and non-standardized product data.

5. Conclusions and Future Work

In this work, we examined the challenges of specificity, level of detail, level of abstraction, and perspective during data exchange in a single domain, on the example of subsurface engineering. Different data models handle the same semantics from different perspectives, which can produce contradictions and incomplete semantic overlap during data integration. Such challenges can be made explicit by the application of the method of reconstruction and communication, which allows the semantic concepts to be mapped to the same common reference frame and compared without distortions stemming from diverging perspectives. The same method can be utilized for a loss- and distortion-free data integration. We demonstrated that the requirements on the data models enabling this process are minimal: the ability to provide containers for holding information, and the ability to label elements and to connect them in arbitrary manner. It is also crucial that such data models are free of any domain-specific semantics, so as not to interfere with the semantic-carrying data models involved in the integration. This makes the clear separation between semantic-carrying and communication-enabling (or integrating) data models necessary, in particular in the development of new data models or for the adaptation of existing ones.

In our future work, we will examine the role of translating data models and the method presented in this paper for the use case of data integration across domain boundaries. Furthermore, we will develop a prototypical framework to test our approach on practical real-

world examples.

Acknowledgement
This work has been supported by the Austrian Research Promotion Agency (FFG) under the GRANT number 886982 (openBAM) and the TransIT (platform for digital transformation in tunneling) project funded by the Federal Ministry of Education, Science and Research under the reference number BMBWF-11.102/0033-IV/8/2019.

References

[1] Hegemann F, Stascheit J, Maidl U and Ninic J 2019 *Proc. of the 14th International Conference Underground Construction, Prague 2019* p 9 URL https://nottingham-repository.worktribe.com/output/2184403
[2] buildingSMART International (bSI) 2022 Industry Foundation Classes accessed 13 January 2022 URL https://technical.buildingsmart.org/standards/ifc/
[3] Beaufils M, Grellet S, Hello B, Lorentz J, Beaudouin M and Castro Moreno J 2019 *Tunnels and Underground Cities: Engineering and Innovation meet Archaeology, Architecture and Art* (Taylor & Francis) pp 655–664 ISBN 9780429424441
[4] Austrian Society for Geomechanics (ASfG) 2021 Richtlinie für die geotechnische Planung von Untertagebauten mit zyklischem Vortrieb accessed 13 January 2022 URL https://www.oegg.eu/en/publications
[5] Rasmussen M H, Lefrançois M, Pauwels P, Hviid C A and Karlshøj J 2019 *Automation in Construction* **108** 16
[6] Swiss Society of Engineers and Architects (SSoEA) 2015 SIA199 - Erfassen des Gebirges im Untertagebau accessed 13 January 2022 URL http://shop.sia.ch/normenwerk/ingenieur/sia%20199/d/2015/D/Product
[7] Eastman C 1975 *AIA Journal* **63** 46–50
[8] Cerovsek T 2011 *Adv. Eng. Informatics* **25** 224–244
[9] Draht R 2010 *Datenaustausch in der Anlagenplanung mit AutomationML: Integration von CAEX, PLCopen XML und COLLADA* (Springer) ISBN 978-3-642-04673-5
[10] Paskaleva G, Wolny S and Bednar T 2018 *Proc. of the International Building Physics Conference (IBPC), Syracuse, NY, USA, September, 2018* pp 1091–1096
[11] Scheider S and Kuhn W 2015 *Applications of Conceptual Spaces: The Case for Geometric Knowledge Representation* (Springer International Publishing) pp 97–122
[12] Mayerhofer T, Wimmer M, Berardinelli L, Mätzler E and Schmidt N 2016 *Proc. of the 4th International Workshop on the Globalization Of Modeling Languages co-located with ACM/IEEE 19th International Conference on Model Driven Engineering Languages and Systems (MODELS 2016), Saint Malo, France, October 4th, 2016* (*CEUR Workshop Proceedings* vol 1731) (CEUR-WS.org) pp 28–37 URL http://ceur-ws.org/Vol-1731/paper_5.pdf
[13] Paskaleva G, Mazak-Huemer A, Waldhart J and Ehrbar H 2021 *Proc. of the 26th IEEE International Conference on Emerging Technologies and Factory Automation, ETFA 2021, Vasteras, Sweden, September 7-10, 2021* (IEEE) pp 1–8
[14] Stark K, Goldschmidt T, Doppelhamer J, Bihani P and Goltz D 2018 *14th IEEE International Conference on Automation Science and Engineering, CASE 2018, Munich, Germany, August 20-24, 2018* (IEEE) pp 645–648
[15] Leng S, Lin J, Li S and Hu Z 2021 *IEEE Access* **9** 148845–148861
[16] Maass W and Lampe M 2007 *37. Jahrestagung der Gesellschaft für Informatik, Informatik trifft Logistik, INFORMATIK 2007, Bremen, Germany, September 24-27, 2007, Band 1* (*LNI* vol P-109) (GI) pp 141–146 URL https://dl.gi.de/20.500.12116/22569