

Extracting JSON Schemas with tagged unions

Stefan Klessinger¹, Meike Klettke², Uta Störl³ and Stefanie Scherzinger¹

¹University of Passau, Passau, Germany

²University of Regensburg, Regensburg, Germany

³University of Hagen, Hagen, Germany

Abstract

With data lakes and schema-free NoSQL document stores, extracting a descriptive schema from JSON data collections is an acute challenge. In this paper, we target the discovery of tagged unions, a JSON Schema design pattern where the value of one property of an object (the tag) conditionally implies subschemas for sibling properties. We formalize these implications as conditional functional dependencies and capture them using the JSON Schema operators `if-then-else`. We further motivate our heuristics to avoid overfitting. Experiments with our prototype implementation are promising, and show that this form of tagged unions can successfully be detected in real-world GeoJSON and TopoJSON datasets. In discussing future work, we outline how our approach can be extended further.

1. Introduction

JSON is a popular data exchange format. Extracting a schema from collections of JSON documents is a real-world challenge which is actively being researched [1, 2, 3, 4, 5, 6, 7, 8].

Ideally, the extracted schema describes the data *tightly*, yet without overfitting. In this article, we target the detection of a specific schema design pattern in the JSON Schema language, the pattern of *tagged unions*, also known as discriminated unions, labeled unions, or variant types. This is a recommended design pattern [9], and has been found to be quite common in real-world schemas [10].

Example 1. Figure 1 shows GeoJSON data. The array starting in line 2 holds four objects, of type `Point` and `LineString`. Point coordinates are encoded as an array of numbers, while lines are encoded as an array of points, and hence, an array of number arrays.

The GeoJSON specification (IETF RFC 7946, clickable link embedded in the PDF) describes six types of geometries, including polygons and multi-polygons. Consistently, the property `type` serves as a *tag* to distinguish the subschema of sibling property `coordinates`, thereby instantiating a tagged union.

GeoJSON comes without an official JSON Schema specification [11]. In Figure 2, we therefore show a hand-crafted excerpt using the `if-then-else` construct to enforce the tagged union, with the following semantics: If the object in question has a property labeled

```
1 { "type": "GeometryCollection",
2   "geometries": [
3     { "type": "Point",
4       "coordinates": [30,10] },
5     { "type": "Point",
6       "coordinates": [40,15] },
7     { "type": "LineString",
8       "coordinates": [[55,5], [10,30], [10,10]] },
9     { "type": "LineString",
10      "coordinates": [[30,10], [10,30], [40,40]] }
11 ] }
```

Figure 1: GeoJSON data. The value of property `type` (here `Point` and `LineString`) determines the subschema of property `coordinates` as either encoding a geometric point (an array of numbers), or a line (an array of points).

`type` with the value `Point`, then the value of property `coordinates` must be an array of numbers. Else, if the property `type` has the value `LineString`, the value of `coordinates` must be an array of points, and hence, an array of number arrays.

Example 2. Various schemas listed on SchemaStore.org, a community-curated collection of JSON Schema declarations, encode tagged unions using `if-then-else`. Among them are the Minecraft schemas which use tagged unions to encode so-called data packs for configuring Minecraft worlds. Other examples are Github Issue Forms, or the cloudify schema (clickable links in the PDF).

In schema extraction, the detection of tagged unions has so far received limited attention: Several approaches are able to detect *union types*, i.e., properties whose type is a union of types, but they do not detect the dependency w.r.t. a specific property value (the “tag” in the tagged union). Figure 3a shows an example: This schema for GeoJSON data does not restrict the values for property `type` (it allows all strings) and allows alternative subschemas for property `coordinates`. As it is overly general, the schema allows to encode lines as mere ar-

DEco@VLDB'22: First International Workshop on Data Ecosystems, September 5, 2022, Sydney, Australia

✉ stefan.klessinger@uni-passau.de (S. Klessinger);

meike.klettke@ur.de (M. Klettke); uta.stoerl@fernuni-hagen.de

(U. Störl); stefanie.scherzinger@uni-passau.de (S. Scherzinger)

🆔 0000-0003-0551-8389 (M. Klettke); 0000-0003-2771-142X

(U. Störl)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

```

1 "if": {
2   "properties": {
3     "type": { "const": "Point" } },
4   "required": [ "type" ] },
5 "then": {
6   "properties": {
7     "coordinates": {
8       "type": "array",
9       "items": { "type": "number" } } } },
10 "else": {
11   "if": {
12     "properties": {
13       "type": { "const": "LineString" } },
14     "required": [ "type" ] },
15   "then": {
16     "properties": {
17       "coordinates": {
18         "type": "array",
19         "items": {
20           "type": "array",
21           "items": { "type": "number" } } } } } } } }

```

Figure 2: JSON Schema snippet declaring a tagged union for GeoJSON geometry objects Point and LineString.

rays of numbers (rather than arrays of number-arrays), in violation of the GeoJSON specification.

Existing approaches to JSON Schema extraction fail to detect tagged unions. Notably, in describing their approach to schema extraction based on typing JSON values, Baazizi et al. [2] outline how their approach can be extended to include tagged unions. However, they do not discuss strategies against overfitting to the input data during the discovery of tagged unions.

Our proposal follows a different approach, and relies on a relational encoding of JSON objects from which we then derive conditional dependencies. A central part of our contribution are our heuristics, which filter out dependencies that have insufficient support in the input data, so that they are not reflected in the derived schema. Specifically, ours is the first proposal towards the discovery of tagged unions that – to our knowledge – includes an experimental evaluation over real-world data.

Contributions. This paper makes the following contributions:

- We target the detection of tagged unions in JSON schema extraction, specifically, tagged unions that are based on dependencies between a property value and the implied subschema of its sibling properties.
- Our approach relies on the discovery of unary constant conditional functional dependencies in a relational encoding of the JSON objects. Traditionally, conditional functional dependencies are employed in the context of data cleaning, and we apply them to a new domain.
- Our approach is composable with existing algorithms for JSON schema extraction, as we impose the tagged unions on top of the schemas derived

by state-of-the-art third-party tools.

- Our experiments on real-world GeoJSON datasets (and its sibling format TopoJSON) show that meaningful tagged unions can indeed be identified. We illustrate the impact of a configurable threshold on the number of tagged unions detected, and consequently, the size of the extracted schema.

We further outline promising directions of future work.

Artifact availability and reproducibility. We have made our research artifacts (code, data, extracted schemas), as well as a fully automated reproduction package, long-term available online [12].

Structure of this paper. In Section 2, we introduce the preliminaries. Section 3 discusses related work. In Section 4, we present our approach, specifically the architecture and heuristics employed, and an outlook on future extensions. In Section 5, our experiments are presented and discussed. In Section 6, we draw our conclusions and discuss opportunities for future work.

2. Preliminaries

2.1. JSON Data Model and JSON Schema

JSON data model.

The grammar below (adopted from [13]) captures the syntax of JSON values, namely basic values, objects, or arrays. Basic values B include the null value, Booleans, numbers n , and strings s . Objects O represent sets of members, each member being a name-value pair, arrays A represent sequences of values.

$$\begin{aligned}
 J &::= B \mid O \mid A \\
 B &::= \text{null} \mid \text{true} \mid \text{false} \mid n \mid s & n \in \text{Num}, s \in \text{Str} \\
 O &::= \{l_1 : J_1, \dots, l_n : J_n\} & n \geq 0, i \neq j \Rightarrow l_i \neq l_j \\
 A &::= [J_1, \dots, J_n] & n \geq 0
 \end{aligned}$$

JSON Schema. JSON Schema is a language for defining the structure of JSON documents. The syntax and semantics of JSON Schema have been formalized in [14], and we informally present some of the main keywords:

Assertions include `required`, `enum`, `const`, `pattern` and `type`, and indicate a test that is performed on the corresponding instance.

Applicators include the Boolean operators `not`, `anyOf`, `allOf`, `oneOf`, as well as, `if-then-else`. They further include the object operators `patternProperties`, `additionalProperties`, and `properties`, and the array operator `items`, and the reference operator `$ref`. They apply a different operator to the same instance or to a component of the current instance.

Annotations (`title`, `description`, and `$comment`) do not affect validation but indicate an annotation associated to the instance.

```

1 "anyOf": [
2   { "type": "object",
3     "properties": {
4       "type": { "type": "string" },
5       "coordinates": {
6         "type": "array",
7         "items": { "type": "number" } } }
8   },
9   { "type": "object",
10    "properties": {
11     "type": { "type": "string" },
12     "coordinates": {
13       "type": "array",
14       "items": { "type": "array",
15                 "items": { "type": "number" } } } } }
16 ] } } ] ]

```

```

1 "anyOf": [
2   { "type": "object",
3     "properties": {
4       "type": { "const": "Point" },
5       "coordinates": {
6         "type": "array",
7         "items": { "type": "number" } } }
8   },
9   { "type": "object",
10    "properties": {
11     "type": { "const": "LineString" },
12     "coordinates": {
13       "type": "array",
14       "items": { "type": "array",
15                 "items": { "type": "number" } } } } }
16 ] } } ] ]

```

(a) Union type.

(b) Tagged union.

Figure 3: Snippets of different encodings of GeoJSON geometries in JSON Schema. Left: Union type encoding, as extracted by traditional schema extraction tools (with syntactic variations, depending on the tool). Right: Tagged union encoding using anyOf, as an alternative to the if-then-else construct. While near-identical in syntax, differing only in lines 4 and 11, the schema semantics differ, as the union type allows incorrect encodings of points or line strings in GeoJSON.

Union types A property with several possible types (or subschemas) can be described as a *union type* [15], i.e., the union of all types the property assumes. In JSON Schema, this can be encoded by a disjunction, using the union operators anyOf or oneOf (where the former is inclusive, and the latter exclusive). Such an encoding is exemplified in Figure 3a for GeoJSON. Union types are recognized by most of the existing tools for JSON Schema extraction (as we will discuss in greater detail in our discussion of related work). While the schema distinguishes two variants for encoding coordinates (an array of numbers, or an array thereof), it does not capture any dependencies between the value of property type and the subschema for the sibling property coordinates. Also, the domain of property type is not restricted to specific string values.

Tagged unions The if-then-else operator allows for declaring tagged unions, and was introduced (rather recently) with JSON Schema Draft 7.

In Example 1, we informally introduced the semantics. As for terminology, we distinguish one property as the *tag* (in our example, the property labeled type), and identify one or more properties with *implied subschemas* (in our example, for coordinates).

The tag-property may go by any name. While the schemas for GitHub Issue Forms and cloudify from Example 2 incidentally also rely on a tag labeled “type”, in the example of Minecraft (see data snippets in Figure 5), the tag-property is labeled “condition”.

An alternative encoding for tagged unions is to use a union operator, as shown in Figure 3b. Two different combinations of values for property type and subschemas for property coordinates are given in accordance with the GeoJSON definition. Again, an object may either be a

“Point” with coordinates encoded as an array of numbers or a “LineString” with an array of points.

Note that while Figures 3a and 3b differ only marginally in their syntax, the difference in semantics is striking: The tagged union captures the dependency between the value of property type and the subschema of property coordinates. While union type encodings can be derived by several state-of-the-art schema extraction tools, schemas with tagged unions are – so far – manually crafted, since existing approaches to schema extraction are not capable of discovering value-based dependencies.

Our approach can produce either the if-then-else encoding or the encoding exemplified in Figure 3b. We chose to implement the former, because it is more lenient w.r.t. unexpected tag values (in this case no restrictions are specified). Note that this is not a limitation of our approach, and a merely technical limitation.

2.2. Dependencies

For the relational model, the concept of functional dependencies is well explored, and various generalizations are known, such as conditional functional dependencies [16] which only apply to a subset of the tuples. In the following, we extend these notions to JSON data, assuming a relational encoding of all objects that are reachable by the same path from the document root. This idea of a relational encoding of semi-structured data for the definition or detection of dependencies is a common approach, e.g., for XML [17], JSON [18], or RDF data [19].

Relational encoding Given a JSON value, we consider all labeled paths from the root to a JSON object. Paths may be encoded in JSONPath [20], a straightforward path language for JSON.

We next introduce the schema for our relational encoding. We reserve attribute $O.id$ for an (internal) object identifier. The object identifier must be unique, but we do not impose any constraints on its semantics. In the following, we will simply use the line of code in the file containing the JSON Schema declaration (after pretty printing), where the scope of the object is first entered.

We identify the labels of all properties reachable by a given path p :

$$L_p = \{l_i \mid \text{object } \{l_1 : J_1, \dots, l_n : J_n\} \\ \text{is reachable by path } p\}$$

For each property label l_i in L_p where J_i occurs as a basic value, we define a relational attribute $l_i.value$ that captures the basic value J_i . These properties are considered to be candidates for tags.

Further, for each property label l_i in L_p , we define an attribute $l_i.type$, capturing the subschema directly derived from its value J_i .

For each object reachable by path p , we then insert one tuple into this relation, choosing some unique object identifier for each object. The attribute values are populated with the semantics described above; *null* values mark missing entries.

Example 3. Table 1 shows the encoding for the JSON objects from Figure 1 in the array starting from line 3, reachable by the JSONPath `/geometries[*]`. By `t1` and `t2`, we abbreviate the subschemas directly derived from the JSON values, requiring an array of numbers (`t1`) and an array of arrays of numbers (`t2`), namely

```
{ "type" : "array",
  "items": { "type": "number" } }
```

 (t1)

and further

```
{ "type": "array", "items": {
  "type": "array",
  "items": { "type": "number" } } }
```

 (t2)

Table 1

Relational encoding for the objects in Figure 1 that are reachable by path `/geometries[*]` (JSONPath syntax). Subschemas `t1` and `t2` abbreviated as in Example 3.

$O.id$	$type.value$	$type.type$	$coordinates.type$
3	Point	string	t1
5	Point	string	t1
7	LineString	string	t2
9	LineString	string	t2

Dependencies We next introduce dependencies over this relational encoding. Traditional functional dependencies (FDs) capture constraints that hold on all tuples of a relation. Moreover, conditional functional dependencies (CFDs) [16] are functional dependencies that hold on only a subset of the tuples. While in their full generality, CFDs are a generalization of classical functional dependencies, we will focus on a very restricted subclass that is related to association rules [21], and that can be defined quite compactly.

Definition 4. If \mathcal{A} is a set of attributes, then a *unary constant conditional functional dependency over \mathcal{A}* (ucCFD) is an expression of the form

$$[A = a] \rightarrow [B = b]$$

where A, B are attributes in \mathcal{A} and a, b are constants from the domains of A and B respectively. A relation R over \mathcal{A} satisfies $[A = a] \rightarrow [B = b]$ if for each pair of tuples $s, t \in R$, $\pi_A(s) = \pi_A(t) = a$ implies $\pi_B(s) = \pi_B(t) = b$.

Example 5. The dependency below holds in Table 1 and reads as follows:

$$[type.value = \text{"Point"}] \rightarrow [coordinates.type = t1]$$

The left-hand-side, in brackets, declares a condition that must be satisfied for the dependency to hold: We consider all attributes where the value of attribute `type.value` is the string “Point”. For all tuples where this condition is satisfied, the value of attribute `coordinates.type` must be the subschema abbreviated as `t1`.

In our domain of application, namely JSON values, these dependencies express powerful constraints between property values and subschemas: In the example dependency above, it is implied that if the value of property `type` is the string constant “Point”, then the value of the sibling property `coordinates` must conform to the subschema `t1`.

In the remainder of this article, we exclusively focus on the discovery of such *value-type* constraints, where the attribute on the left-hand-side of a ucCFD is of the form “ $A.value$ ” (the value of property A , which we consider to be a candidate for a tag in a tagged union), and the attribute on the right-hand-side is of the form “ $B.type$ ” (the subschema of the sibling property B).

Traditionally, tagged unions are declared by switching on the value of a *single* tag property. This is also the recommended practice in JSON Schema [9], and in agreement with what we observe in real-world data [10]. We are therefore confident that our restriction to unary dependencies is justified.

3. Related Work

Our article builds upon the rich body of related work in the area of schema extraction and the theory of data dependencies.

Schema and constraint definition. Schema languages for semi-structured data are well-researched.

XML was developed as a semi-structured data format with implicit structural information and an optional explicit schema. The simplest schema language for XML is DTD (Document Type Definition). The lack of means to express data types and exact cardinalities in DTDs motivated the development of further schema languages, such as XML Schema [22], Schematron [23], and RelaxNG [24]. All three support the definition of constraints. In XML Schema, alternatives can be defined by specifying conditions on path expressions (XPath). In version 1.1 of XML Schema, the concept of assertions allows to encode constructs such as tagged unions. Schematron can define rules with context information (XPath) and messages (that are sent in the success or error case). RelaxNG enables the definition of inference rules. This historical excursion shows the necessity of exact and expressive schema languages. The same holds true for the description of JSON data. The preliminaries on JSON Schema were already covered in paragraph 2.1.

Schema extraction. In *relational databases* which follow a schema-first approach, all available databases have an explicit schema stored in the databases catalog. With semi-structured data, on the other hand, there are many datasets available that have been published without such explicit schema information. Schema extraction (also known as reverse engineering) is therefore an important subtask in data profiling for semi-structured data.

XML schema extraction. For schema extraction from XML documents, different approaches have been developed, (e.g. [25, 26, 27, 28, 29, 30]). In all algorithms, a simple schema consisting of element and attribute names, nesting, optional and required information is derived. We are not aware of any approaches to extracting complex schema constraints, such as tagged unions.

JSON schema extraction. Early work on schema extraction from JSON data [1] adds – besides the schema itself – also the extraction of statistics and a detection of outliers. In [6], the extraction of schema versions over time, as well as evolution operations mapping between consecutive schema versions, is presented.

Recent surveys of different schema extraction approaches were provided by [4] (qualitative comparison) and [5]

(quantitative comparison). Several of the examined approaches also support the extraction of union types: [1] and [31] use the JSON Schema keyword `oneOf`, while [2] use the union type constructor of their own proprietary schema language. The authors of [32] encode the extracted schema in the XML Schema language, and encode union types using entity versioning, while [33] pursue an alternative approach of reducing different types to their most generic type.

But neither Klettke et al. [1] nor Frozza et al. [31] support tagged unions within their extraction of union types. The approach by Baazizi et al. [2] is based on type inference. They achieve scalability by inferring types in a MapReduce-based approach. The authors do discuss the challenge of extracting tagged unions, and describe an extension to their algorithm to address this challenge. However, there is no implementation or evaluation of this feature, or of any heuristics to prevent overfitting in tagged unions.

In a further, recent contribution, Spoth et al. [3] focus on resolving ambiguities during schema extraction, such as sets encoded as objects rather than arrays. As their approach does not consider property values, it cannot detect tagged unions.

Durner et al. [34] recently presented an approach for fast analytics on semi-structured data. Using various algorithms, the JSON data is divided into *tiles*, and local schemas are extracted. However, tagged unions are neither considered nor recognized.

A completely different approach for schema inference has been suggested in [35]. In this work, a supervised learning method (based on the well-known C4.5 classification algorithm) is used for detecting hidden rules in the different variants of datasets. These rules can be either structure-based or value-based. An interesting observation of this work was that based on an empirical study (interviews), value-based rules are considered more important by human consumers than structure-based rules to distinguish variants. The result of the approach is a decision tree that distinguishes the different schema variants with value-based or structure-based conditions on each edge. Even if the approach in the paper is very different from ours, we recognize the motivation that considering value-based conditions in heterogeneous databases is important.

Dependencies. For an overview of data dependencies, we refer to a comprehensive survey [36], and focus on functional, inclusion, and conditional functional dependencies in the following.

Functional dependencies and inclusion dependencies. Functional dependencies [37, 38, 39, 18] and inclusion dependencies on semi-structured data [40, 18]

define semantic constraints on data to guarantee certain data characteristics, to normalize data [41], and to ensure data quality. These constraints can also be applied in data cleaning. In [42], an overview over different methods and the semantic constraints, rules, or patterns to detect errors in the data cleaning process is given. Schelter et al. [43] use declarative rules in unit tests to check the data quality (defined with several metrics). Some examples are: completeness of datasets, range conditions, certain cardinalities and constraints on statistics. Both approaches do not infer the semantic constraints from data but show how declarative rules can be leveraged during data preprocessing.

The discovery of valid uniqueness constraints, functional dependencies and inclusion dependencies is a well-studied field in relational databases [44, 45, 46, 47, 48] and even more relevant in JSON data (or, more generally, NoSQL data) because often these semantic constraints are not predefined in the NoSQL databases. They are implicitly available in the data but for data cleaning tasks, this information is required in form of rules and constraints. Hence, development of algorithms to derive explicit semantic constraints from NoSQL data is of particular importance.

Methods for deriving semantic constraints from data can build upon the algorithms developed for relational databases. Additionally, the algorithms have to scale with large volumes of data, and be robust despite the heterogeneity of datasets (variety) and low data quality (outliers in datasets).

Arenas and Libkin define functional and inclusion dependencies for XML [17], as a basis for schema normalization. More recently, Mior [18] targets the mining of functional and inclusion dependencies from JSON data. Both formalisms do not allow to capture the conditional functional dependencies required in our context.

However, Mior compares the performance of dependency discovery in a relational encoding of the JSON data (termed “static unrolling”) with a dynamic unrolling technique, showing that the latter has superior runtime performance. Mior focuses on the discovery of functional dependencies and inclusion dependencies. While his approach does not consider the special case of conditional functional dependencies, which is relevant in our context, it is capable of deriving approximate dependencies, thus being more robust w.r.t. vagueness in the input data.

Kruse et al. developed in [40] the scalable discovery of inclusion dependencies from data. Scalability of the approach is achieved by two reasons: the approach is only concentrating on discovery of unary inclusion dependencies (consisting of one attribute on the left-hand and on the right-hand side) and is developing a distributed approach. Combining both, a very efficient algorithm can be developed for analyzing large datasets. The heterogeneity of datasets has not been considered in this

approach. In previous work on inferring inclusion dependencies from JSON data [6], we suggested an efficient method that uses lattice characteristics to optimize the algorithm. To consider outliers, a threshold is introduced. With it, it is possible to derive related inclusion dependencies which are violated by a small number of outliers.

A further relaxation of functional dependency discovery has been suggested in [49] for data lakes. Here, outliers are not ignored, but properties that have different yet similar labels are combined. This method can be used for data exploration and profiling of datasets with lower data quality.

Conditional functional dependencies. Conditional functional dependencies [16] were first introduced for relational data, in the context of data cleaning. We apply conditional functional dependencies in a new context, the relational encoding of JSON objects.

In [50, 51], three algorithms for CFD mining are proposed and evaluated: CTANE and FastCFD build upon existing algorithms for FD discovery, TANE and FastFD, respectively and are intended for general CFD discovery. The third algorithm is custom-designed for constant CFDs, as also targeted by us.

In [52], additional rules for pruning the search space and consequently speeding up the mining process are proposed for constant CFDs. Further approaches, based on FD discovery and pattern mining, are explored in [21].

4. Approach

Our end-to-end approach is sketched in Figure 4. In our upcoming walk-through of the architecture, we first focus on the basic approach, as well as our heuristics against overfitting and schema bloat. We then outline future work extensions to robustly generalize our approach to a larger family of tagged unions.

4.1. Architecture

Composability. We design our approach to be composable with existing schema extraction algorithms: Given a collection of JSON values, we extract a JSON Schema description S , using any third-party tool. Along the process sketched along the bottom, we further derive the declaration of tagged unions, denoted T , a sequence of (possibly nested) JSON Schema `if-then-else` statements.

By the conjunction of the subschemas as $\{ \text{"allOf"} : [S, T] \}$, the composite schema requires the JSON values to satisfy both subschemas, imposing tagged unions on schema S . This approach allows us to focus on the particular challenge of identifying tagged unions, while conveniently leveraging state-of-the-art schema extraction tools for describing the other aspects of the

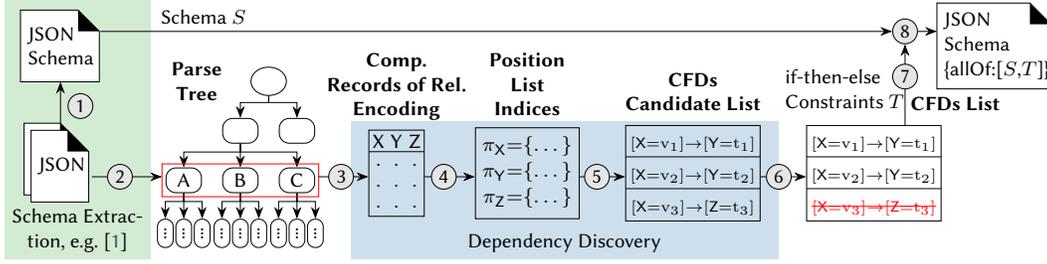


Figure 4: System architecture overview. (1) A third-party tool is used to extract a JSON Schema description S of JSON input data. Steps (2) through (7) visualize the discovery of tagged unions as schema T , and are described in Section 4. In step (8), the schemas S and T are composed into a composite schema. Boxed areas capture state-of-the-art algorithms integrated in our architecture. A, B and C are JSON property labels while X, Y and Z are attributes in our relation encoding (e.g., A.value).

schema. We leave it to future work to recursively merge the composite schema, thereby improving the readability as well as the locality of the interacting constraints.

Relational encoding. In a first step towards the discovery of tagged unions, we parse all JSON values into a single parse tree (adding a virtual root node), and determine the set of all labeled paths from the root to a JSON object. For each such path, we then derive the relational encoding of all properties in the reachable objects. We have already introduced this data structure in Section 2, and resume the discussion of our running example.

Example 6. Consider the GeoJSON value in Figure 1. Table 1 shows the relational encoding of the four objects reachable by the path `/geometries[*]`. The role of the unique object identifier is to preserve cardinalities in repeating attribute values, allowing us to later filter dependencies based on their support in the input.

The running example has been designed to be simple, and it ignores challenges such as properties with varying types, dealing with several properties whose values have basic types, or null-values in the relational encoding. Our approach can be extended to robustly handle these cases as well, and we refer to Section 4.3 for a discussion on the most interesting future generalizations.

Dependency discovery. Based on the relational encoding, we discover conditional functional dependencies. We restrict ourselves to unary constant conditional dependencies in our relational encoding, of the form

$$[A.value = c] \rightarrow [B.type = \sigma]$$

where the left-hand-side denotes the value of the candidate tag, and the right-hand-side a property with an implied subschema. Above, A and B are distinct property labels, c is a basic-value constant, and σ denotes the

implied subschema. Recall that with the dot-notation, we distinguish the value of JSON property A (i.e., $A.value$) from the subschema of sibling property B (i.e., $B.type$) in our relational encoding of JSON objects.

Example 7. From the relational encoding in Table 1, we derive the following dependencies, again using the abbreviations $t1$ and $t2$ for the subschemas introduced earlier:

$$\begin{aligned} [\text{type.value} = \text{"Point"}] &\rightarrow [\text{coordinates.type} = t1] \\ [\text{type.value} = \text{"LineString"}] &\rightarrow [\text{coordinates.type} = t2] \end{aligned}$$

Recall that in the context of our application, they are interpreted as follows. Given a JSON object reachable by the path `/geometries[*]`, if the value of property `type` is “Point”, the value of the sibling property `coordinates` must conform to the subschema $t1$ (an array of numbers). If the value of property `type` is “LineString”, then the value of sibling property `coordinates` must conform to the subschema $t2$ (an array of points).

In the discovery of these dependencies, we leverage state-of-the-art algorithms [53, 47, 54, 55] (originally developed for the discovery of traditional functional dependencies), such as a compressed-records encoding of the relation, computing position list indexes (PLIs) and producing a sequence of candidate dependencies. PLIs are calculated for each column of the relational encoding and allow to effectively group all rows (identified by their row number) with the same value in the same subset.

In the following, we outline the general idea. Inferring conditional functional dependencies from PLIs can be achieved by finding inclusions between them: given PLIs $\pi_{A.value} = \{\{1, 2, 3\}, \{4, 5, 6\}\}$ of attribute $A.value$ and $\pi_{B.type} = \{\{1, 2, 3, 4\}, \{5\}, \{6\}\}$ of attribute $B.type$, we receive the inclusion $\{1, 2, 3\} \subset \{1, 2, 3, 4\}$, allowing us to infer that the values of JSON property A corresponding to rows 1, 2 and 3 determine the subschemas of the sibling property B .

More sophisticated algorithms developed for the discovery of conditional functional dependencies (CFDs),

such as CFDMiner, CTANE and FastCFD [50], could be deployed just as well. As our current restrictions, most notably the one to unary dependencies, obfuscate many of the challenges in FD and CFD discovery, we opted for our simpler approach.

Pruning candidate dependencies. To avoid overfitting in the derived schemas, we heuristically prune candidate dependencies, as described in the upcoming Section 4.2.

Finalization. In a final step, the conditional functional dependencies are transformed into nested `if-then-else` statements in JSON Schema T . This step is merely technical, as is the construction of the composite schema.

4.2. Heuristics

In our approach, an inherent challenge is dealing with high numbers of discovered dependencies, which may lead to overfitting the derived schema w.r.t. the input data. We next distinguish default heuristics and heuristics that are configurable by an expert user.

4.2.1. Default Heuristics

The following heuristics are always applied, as they prevent schema bloat due to overfitting.

Single-valued attributes. In discovering dependencies, we ignore all attributes from the relational encoding with a single-valued domain. In Table 1, this concerns the attribute `type.type`.

Unique attributes. Unique property values are recognized as conditional functional dependencies, and ultimately, cause overfitting and schema bloat. In dependency discovery, we therefore ignore all attributes where the domain consists of unique values (such as the object identifier `O.id`, which does not appear in the input data, and need not be considered in dependency discovery).

Union rule. We apply the union rule to the discovered dependencies, so that one tag may imply the subschemas for several sibling properties. This improves schema succinctness.

4.2.2. Configurable Heuristics

The following heuristic is reasonable in many cases. As it is not universally applicable, we make it configurable.

Minimum Threshold π_{min} . Especially in large datasets, it is to be expected that dependencies are inferred that have no real semantics, despite us ignoring attributes with unique values. Our implementation has a configurable *minimum threshold*: Dependencies with insufficient support in the input are then ignored during the generation of `if-then-else` statements.

By allowing to configure this threshold, either as an absolute or relative value, the number of discovered dependencies can be influenced. Our experiments evaluate this effect.

4.3. Future Generalizations of the Approach

In GeoJSON data, which served as our motivation for the discovery of tagged unions, the geometry objects are highly regular. When inspecting further instances of tagged unions in real-world data, we encounter variations that require us to generalize our approach. For illustration, we again resort to an example.

```
1 [ { "condition": "minecraft:time_check",
2     "period": 24000,
3     "value": { "min": 0.0 } } ]
```

(a) The value of property `min` (line 3) is numeric.

```
1 [ { "condition": "minecraft:weather_check",
2     "raining": false,
3     "thundering": false },
4   { "condition": "minecraft:time_check",
5     "period": 24000,
6     "value": {
7       "min": {
8         "target": {
9           "name": "%StartSunburn",
10          "type": "minecraft:fixed" },
11          "score": "Daytime",
12          "type": "minecraft:score" } } } ]
```

(b) The value of property `min` (line 7) is of type object.

Figure 5: Minecraft JSON data (snippets edited and shortened, clickable links to GitHub source in the subcaptions). Property `condition` functions as the tag in the tagged union.

Example 8. We consider JSON data used in the Minecraft game, as shown in Figure 5. The property `condition` functions as a tag, and distinguishes notions such as “time checks” and “weather checks”, among others. Depending on the tag value, different sibling properties exist: Given that the tag value is a “weather check”, sibling properties labeled `raining` and `thundering` exist. Given that the tag value is a “time check”, sibling properties labeled `period` and `value` exist.

To extend our approach to such more general forms of tagged unions, we need to be able to detect conditional dependencies between the value of a candidate tag,

and the existence of a sibling property (which we term *value-label* constraints, to distinguish from the value-type constraints that we already handle). Here, a specific challenge is to handle optional sibling properties, which are not always present.

Example 9. Resuming the previous example of a Minecraft “time check”, the subschema for property `value` is an object with a property `min`. The subschema of property `min`, however, may vary. In the top example, it is a numeric value. In the bottom example, it is a nested object. Our approach outlined so far cannot detect this conditional functional dependency, because the subschemas of the sibling property differ.

To detect value-type dependencies where the types on the right-hand-side may vary, we need to relax the recognized subschema. In our current prototype implementation, we introduce attributes $\text{min.type}_1, \dots, \text{min.type}_k$ (e.g., for a fixed limit $k = 6$), where we describe the subschema with a varying degree of detail. For instance, from a shallow $\text{min.type}_1 = \{ \text{"type": "object"} \}$ to a very fine-grained description that fully captures all nested objects. In pruning candidate dependencies, we only keep the most detailed subschema for which a CFD still holds. This allows us to detect a common subschema in many practical cases.

5. Experiments

We next describe our experiments using real-world geospatial datasets encoded in JSON. For notational simplicity, we will write “CFDs” when we refer to the subfamily of unary constant conditional functional dependencies, as introduced in Section 2.

5.1. Setup

Implementation. Our prototype is implemented in Python, using the library `anytree` for managing parse trees. For schema validation, we use the Python library `jsonschema` (clickable links in the PDF).

For the “third-party” schema extraction tool, we employ a tool that we have built in earlier work [1]. Note that our approach is designed to work with any other tool for JSON Schema extraction. Also, since we construct the composite schema by combining two subschemas, we are not restricted to Python-based tools.

To confirm the composability of our approach, we have successfully built a variant of our architecture based on the schema extraction tool by Frozza et al. [31, 56] (with minor technical adaptations), but we do not report the results, since they do not contribute new insights.

Artifact availability and reproducibility. Our research artifacts (code, data, extracted schemas), as well

as a fully automated reproduction package, are available online [12]. Additionally, in the PDF version of this article, clickable links in Table 2 allow our readers to directly inspect the input datasets, as well as the extracted schemas.

Datasets. We consider the five JSON datasets listed in Table 2, all in the GeoJSON or the related TopoJSON format. For simplicity, we will commonly treat the formats GeoJSON and TopoJSON synonymously, in discussing our results.

We identified these datasets by manually searching open data servers, as well as by performing a search over all open-source licensed repositories on GitHub using Google BigQuery. Our selection criteria were (1) a sufficiently large document size and (2) that at least two different types of geometries co-occur, so that we may be able to infer tagged unions.

We briefly describe the datasets:

- *Map Germany* is a TopoJSON dataset, describing a map of Germany and surrounding areas, including cities, urban areas, and rivers. This dataset uses geometry types Point, LineString, MultiLineString, Polygon and MultiPolygon.
- *Map EU* is a TopoJSON dataset. Its structure is comparable to *Map Germany* but describes countries in the EU in detail and contains the borders of some adjacent countries.
- *WorldGen Berlin* is a GeoJSON dataset from OpenStreetMap.
- *US House Members* lists members of the US House of representatives, including their place of birth, hometown and congressional district, encoded in GeoJSON.
- *Wildlife Sites* contains wildlife sites in West Yorkshire, described either as a GeoJSON Polygon or MultiPolygon.

The datasets encode GeoJSON or the related TopoJSON format (distinguished as formats “T” and “G” in the table). We list the size of each dataset in lines of code after pretty-printing (column $|D|$).

In lieu of a ground truth describing the number of conditional functional dependencies in the data that is not GeoJSON-encoded, we report in the table the number of detectable CFDs in the GeoJSON-parts only (Det_{Geo}). We manually determine this value for each dataset, inspecting all distinct types of GeoJSON objects (e.g., Point, LineString, Polygon) on each different path. We ignore all paths where the value of the GeoJSON type does not vary. While these can be detected as CFDs from the data, they are not meaningful for recognizing tagged unions, and should be modeled, for instance, simply as constant values instead.

Table 2

Summary of experimental results. Links embedded in the PDF refer to the original dataset (D). We state the dataset format (TopoJSON or GeoJSON) and report its size ($|D|$) in lines of code (LoC), the size of the conventionally extracted schema S (in LoC) and the number of theoretically detectable Geo-/TopoJSON CFDs (Det_{Geo}). For different threshold settings π_{min} , we produce a subschema T encoding the tagged unions. “Ratio T ” reports the share of T w.r.t. the size of the composite schema. Further, the number of CFDs initially discovered (w/o heuristics), after applying threshold π_{min} (see Section 4.2.2), and after applying the remaining default heuristics (see Section 4.2.1) are reported. The last column reports the result of checking whether dataset D is valid w.r.t. the extracted composite schema (a checkmark symbol “✓” denotes a successful check). The checkmarks are clickable links, embedded in the PDF, to the composite schema.

Dataset D	Format	$ D $ (LoC)	$ S $ (LoC)	Det_{Geo}	π_{min}	$ T $ (LoC)	Ratio T	CFDs w/o Heuristics	CFDs w/ π_{min}	CFDs w/ π_{min} and Heuristics	Valid
Map Germany	T	181 585	405	6	50%	126	23.6%		11	4	✓
					35%	158	27.9%	44	14	5	✓
					15%	190	31.7%		20	6	✓
Map EU	T	122 380	359	6	50%	95	20.7%		6	3	✓
					35%	95	20.7%	10	6	3	✓
					15%	127	25.9%		8	4	✓
WorldGen B.	G	56 441	1 339	11	50%	150	10.0%		76	4	✓
					35%	182	11.9%	1 243	109	5	✓
					15%	278	17.1%		184	8	✓
US House M.	G	22 745	493	2	50%	54	9.8%		459	2	✓
					35%	54	9.8%	719	470	2	✓
					15%	54	9.8%		660	2	✓
Wildlife Sites	G	876 482	143	2	50%	38	20.5%		18	1	✓
					35%	38	20.5%	360	18	1	✓
					15%	92	38.5%		50	3	✓

Execution environment. Our experiments were conducted on an off-the-shelf notebook with an Intel i7-1165G7 CPU with 4.7GHz and 16GB of main memory.

We use Det_{Geo} as a target in discussing recall of dependency discovery in the GeoJSON part of the input.

5.2. Experimental Design

Workflow. For each data collection, we extract composite schemas choosing three settings for the *minimum threshold* π_{min} : an aggressive setting of 50%, requiring a conditional functional dependency to occur in at least half of the tuples in the relational encoding of all objects reachable by a given path, and more lenient settings with 35% and 15%, respectively.

For each threshold setting, we generate schemas S (from the third-party tool) and T (encoding the tagged unions), as well as the composite schema.

Each composite schema is validated against the specification of JSON Schema Draft-07, confirming that the composite schema conforms. The input datasets are further validated against the extracted composite schema, checking for any logical errors in schema extraction.

Metrics. In our quantitative assessment, we report schema sizes (after syntactic normalization by pretty-printing), the number of conditional functional dependencies initially discovered, and after filtering according to our heuristics.

5.3. Results

Table 2 summarizes our results. Note that all composite schemas successfully validate against Draft-07, and further, that all JSON datasets validate against the composite schemas.

Decreasing the threshold π_{min} leads to more conditional functional dependencies being discovered. The share of the subschema encoding tagged unions compared to the entire composite schema ranges between 10% and up to approx. 38%. While the 38% share might appear to be large, the composite schema has less than 250 lines in total, compared to 800K lines of JSON input. Thus, this dataset is highly regular in its structure, and the extracted schema comparatively compact.

For the bottom three datasets, we detect conditional functional dependencies in the hundreds, even more than one thousand in the case of the WorldGen Berlin data. However, applying the threshold and the other heuristics drastically reduces the dependencies towards the number of tagged unions that we expect to find (c.f. the target Det_{Geo}). Note that there are only six distinct geometries in GeoJSON which may, however, occur on different paths, leading to more than six if-then-else

statements in the GeoJSON part.

After applying heuristics and a threshold of 15%, we only miss two detectable CFDs in Map EU and three in WorldGen Berlin. Missed CFDs can be attributed to the threshold being too coarse.

Manually inspecting the final dependencies, we find one unexpected dependency for Wildlife Sites with a threshold of 15%. This is the only false positive dependency across all datasets. It results in a tagged union being declared for a specific date, which is not semantically meaningful. Further, this dependency is outside the GeoJSON-encoded part of the data, explaining why the number of detected CFDs is higher than Det_{Geo} . Thus, we have a single case of overfitting for the datasets analyzed. Overall, we observe very high precision in recognizing the GeoJSON/TopoJSON dependencies resident in the data.

We do not focus on runtime evaluation, as our implementation is prototypical and unoptimized. Yet to provide a general perspective, we share our observation that runtimes vary greatly between datasets, ranging from roughly one second to approximately one minute. This is a waiting time which we deem acceptable for a non-interactive, irregular task.

Since our approach is main-memory based, memory is a physical limitation to the inputs that we can process. For the datasets considered here, 16GB of RAM are sufficient to run our experiments.

5.4. Discussion

The comparison between the number of dependencies found with different heuristics shows the effectiveness of these heuristics in pruning dependencies. In particular, this concerns the configuration knob represented by *minimum threshold*.

Our approach to setting this threshold is rather coarse, with 50% obviously too rigid, but 15% delivering meaningful results. In future work, we plan to investigate how to auto-adjust the threshold, based on statistical distributions obtained by profiling the input.

An obvious threat to the generalizability of our results – however promising they are – is that the GeoJSON datasets are highly regular in their structure. In the GeoJSON-part, they contain only a small number of detectable dependencies, between two and eleven. Therefore, we refrain from computing the metrics precision and recall, as they are easily distorted when working with small numbers. However, the manual inspection of the derived dependencies, and the comparison against the target Det_{Geo} , shows that our approach is indeed successful for the input datasets chosen.

To counter the threat of generalizability, further experiments over different real-world datasets (e.g., the datasets

listed in Example 2), as well as synthetic data, and data not encoding any tagged unions, are required.

We do not focus on the performance of our unoptimized prototype implementation, as the algorithm is only executed once for each dataset. Of course, scalability to larger inputs is an issue to be addressed in future work, as CFD discovery in general is an expensive problem: Already in addressing the discovery of functional dependencies in their full generality, algorithms have exponential runtime complexity [57]. Several algorithms for the specific problem of CFD discovery have been proposed [50]. Their evaluations show that runtime performance depends heavily on the input, with some algorithms scaling better with size of the dataset (i.e., the number of tuples) thus being suitable for large datasets, while others perform better with higher arity. For constant CFDs, Li et al. [52] achieved promising runtime improvements by applying custom rules for pruning the search space.

With strong restrictions to the problem space, such as our restriction to unary constant CFDs in our case, we can expect reasonable runtime performance on real-world inputs (which are often reasonably well-behaved).

In our experiments, we observed a considerable impact of our heuristics and threshold on runtime, reducing the processing time for all datasets by up to two orders of magnitude. This is promising for further optimizations.

6. Conclusion and Outlook

In this article, we proposed a method to infer conditional functional dependencies from JSON data. We use the identified dependencies as the basis for declaring tagged unions in JSON schema extraction.

This allows us to capture value-based constraints. In fact, in [35] Gallinucci et al. report on expert interviews with users, regarding the users' preferences in schemas over nested data. Their interviews reveal that value-based conditions have a greater influence on the differentiation of schema variants than structural constraints, and are therefore preferred.

In future work, our approach can be extended for the extraction of further variants of dependencies, for instance, traditional conditional functional dependencies that capture implications between atomic property values, or dependencies where the value of the tag property implies the existence of a specific sibling property, as discussed in greater detail in Section 4.3. This allows to recognize a larger family of tagged unions.

Our prototype implementation is currently main-memory based, which limits the size of inputs that we can handle. Making our implementation scale to larger inputs is one of the immediate next steps. Here, we may build upon first results by Mior [18], who shows that discover-

ing dependencies in a relational encoding of JSON data has inferior runtime performance when compared to discovering dependencies in a streaming fashion. Also, we plan to consider a MapReduce-based approach to schema extraction, as implemented by Baazizi et al. in [2].

Ultimately, our goal is to obtain a schema declaration that human consumers consider to be comprehensive, but that may also be efficiently processed programmatically. In order to obtain more succinct schemas, we need to resolve redundancies between the schema extracted by a third-party tool and our encoding of tagged unions. This requires rewriting the composite schema based on an algebraic representation of JSON Schema operators, such as the schema algebra proposed by Attouche et al. [13].

Further, identifying metrics that capture the quality of the extracted schemas will allow to quantitatively compare schemas extracted by different approaches. Given suitable metrics, the configuration of heuristics could even be adjusted automatically. A possible direction is to explore the notions of precision and recall, and the proxy-metric of schema entropy, as introduced by Spoth et al. in [3].

A further task is to adopt a CFD inference algorithm that is robust despite poor data quality and that can infer constraints despite outliers in the data. Naturally, this requires a relaxation to “soft” CFDs, a task where we may also build upon existing work on relational [21] and even JSON data [18].

In summary, our long-term vision is to extract comprehensible and therefore human-consumable schema declarations from JSON data. We believe that the detection of schema design patterns that are popular among schema designers, such as tagged unions, is an important building block towards realizing this vision.

Acknowledgments: This work was funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) grant #385808805. We thank Thomas Kirz for expertly typesetting the systems architecture in \LaTeX . We thank the anonymous reviewers of DEco’22, especially reviewer #2, for the detailed and helpful feedback.

References

- [1] M. Klettke, U. Störl, S. Scherzinger, Schema extraction and structural outlier detection for JSON-based NoSQL data stores, in: Proc. BTW, volume P-241, GI, 2015, pp. 425–444.
- [2] M. A. Baazizi, D. Colazzo, G. Ghelli, C. Sartiani, Parametric schema inference for massive JSON datasets, VLDB J. 28 (2019) 497–521.
- [3] W. Spoth, O. Kennedy, Y. Lu, B. C. Hammerschmidt, Z. H. Liu, Reducing Ambiguity in JSON Schema Discovery, in: Proc. SIGMOD, 2021, pp. 1732–1744.
- [4] P. Contos, M. Svoboda, JSON schema inference approaches, in: Proc. ER (Workshops), 2020, pp. 173–183.
- [5] I. Veinhardt Latták, P. Koupil, A comparative analysis of JSON schema inference algorithms, in: Proc. ENASE, 2022, pp. 379–386.
- [6] M. Klettke, H. Awolin, U. Störl, D. Müller, S. Scherzinger, Uncovering the evolution history of data lakes, in: Proc. Big Data, 2017, pp. 2462–2471.
- [7] M. A. Baazizi, C. Berti, D. Colazzo, G. Ghelli, C. Sartiani, Human-in-the-loop schema inference for massive JSON datasets, in: Proc. EDBT, 2020, pp. 635–638.
- [8] J. Namba, Enhancing JSON schema discovery by uncovering hidden data, in: Proc. VLDB 2021 PhD Workshop, 2021.
- [9] M. Droettboom, Understanding JSON schema, <https://json-schema.org/understanding-json-schema/reference/conditionals.html#if-then-else>, 2022. Draft 2020-12.
- [10] M. A. Baazizi, D. Colazzo, G. Ghelli, C. Sartiani, S. Scherzinger, An empirical study on the “usage of not” in real-world JSON schema documents, in: Proc. ER, 2021, pp. 102–112.
- [11] A. A. Frozza, R. dos Santos Mello, Js4geo: A canonical JSON schema for geographic data suitable to NoSQL databases, Geoinformatica 24 (2020) 987–1019.
- [12] S. Klessinger, M. Klettke, U. Störl, S. Scherzinger, Extracting JSON Schemas with Tagged Unions (Reproduction Package), 2022. URL: <https://doi.org/10.5281/zenodo.6985647>. doi:10.5281/zenodo.6985647.
- [13] L. Attouche, M. A. Baazizi, D. Colazzo, G. Ghelli, C. Sartiani, S. Scherzinger, Witness generation for JSON schema, CoRR abs/2202.12849 (2022). URL: <https://arxiv.org/abs/2202.12849>. arXiv:2202.12849.
- [14] F. Pezoa, J. L. Reutter, F. Suárez, M. Ugarte, D. Vrgoc, Foundations of JSON Schema, in: Proc. WWW, 2016, pp. 263–273.
- [15] F. Suárez, J. Reutter, D. Vrgoc, M. Ugarte, F. Pezoa, JSON Schema: Multiple Types (Software), https://cswr.github.io/JsonSchema/spec/multiple_types/, 2016.
- [16] P. Bohannon, W. Fan, F. Geerts, X. Jia, A. Kementsietsidis, Conditional functional dependencies for data cleaning, in: Proc. ICDE, 2007, pp. 746–755.
- [17] M. Arenas, L. Libkin, A normal form for xml documents, ACM Trans. Database Syst. 29 (2004) 195–232.
- [18] M. J. Mior, Fast discovery of nested dependencies on JSON data, CoRR abs/2111.10398 (2021). URL: <https://arxiv.org/abs/2111.10398>. arXiv:2111.10398.
- [19] S. Kruse, A. Jentsch, T. Papenbrock, Z. Kaoudi,

- J. Quiané-Ruiz, F. Naumann, Rdfind: Scalable conditional inclusion dependency discovery in RDF datasets, in: Proc. SIGMOD, 2016, pp. 953–967.
- [20] J. Friesen, Java XML and JSON: Document Processing for Java SE, Apress, 2019, pp. 299–322.
- [21] J. Rammelaere, F. Geerts, Revisiting conditional functional dependency discovery: Splitting the "C" from the "FD", in: Proc. ECML PKDD, 2018, pp. 552–568.
- [22] W3C, W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures, W3C Recommendation, 2012. URL: <https://www.w3.org/TR/xmlschema11-1/>.
- [23] ISO/IEC, ISO/IEC 19757-3:2020 Information technology – Document Schema Definition Languages (DSDL) – Part 3: Rule-based validation – Schematron, ISO/IEC Standard, 2020. URL: <https://www.iso.org/standard/74515.html>.
- [24] ISO/IEC, ISO/IEC 19757-2:2008 Information technology – Document Schema Definition Language (DSDL) – Part 2: Regular-grammar-based validation – RELAX NG, ISO/IEC Standard, 2008. URL: <https://www.iso.org/standard/52348.html>.
- [25] M. Garofalakis, A. Gionis, R. Rastogi, S. Seshadri, K. Shim, XTRACT: A system for extracting document type descriptors from XML documents, in: Proc. SIGMOD, 2000, pp. 165–176.
- [26] C.-H. Moh, E.-P. Lim, W.-K. Ng, DTD-Miner: a tool for mining DTD from XML documents, in: Proc. WECWIS, 2000, pp. 144–151.
- [27] J. Hegewald, F. Naumann, M. Weis, Xstruct: Efficient schema extraction from multiple and large XML documents, in: Proc. Workshops ICDE, 2006.
- [28] M. Klempa, M. Kozák, M. Mikula, R. Smetana, J. Stárka, M. Svirec, M. Vitásek, M. Necaský, I. Holubová, *JInfer*: A framework for XML schema inference, Comput. J. 58 (2015) 134–156.
- [29] B. Chidlovskii, Schema extraction from xml collections, in: Proc. JCDL, 2002, p. 291–292.
- [30] I. Mlýnková, M. Nečaský, Heuristic methods for inference of xml schemas: Lessons learned and open issues, Informatica 24 (2013) 577–602.
- [31] A. A. Frozza, R. dos Santos Mello, F. de Souza da Costa, An approach for schema extraction of JSON and extended JSON document collections, in: Proc. IRI, 2018, pp. 356–363.
- [32] D. S. Ruiz, S. F. Morales, J. G. Molina, Inferring versioned schemas from NoSQL databases and its applications, in: Proc. ER, 2015, pp. 467–480.
- [33] J. L. C. Izquierdo, J. Cabot, Discovering implicit schemas in JSON data, in: Proc. ICWE, 2013, pp. 68–83.
- [34] D. Durner, V. Leis, T. Neumann, JSON tiles: Fast analytics on semi-structured data, in: Proc. SIGMOD, 2021, pp. 445–458.
- [35] E. Gallinucci, M. Golfarelli, S. Rizzi, Schema profiling of document-oriented databases, Inf. Syst. 75 (2018) 13–25.
- [36] S. Song, F. Gao, R. Huang, C. Wang, Data dependencies extended for variety and veracity: A family tree, IEEE Transactions on Knowledge and Data Engineering 34 (2022) 4717–4736.
- [37] M. L. Lee, T. W. Ling, W. L. Low, Designing functional dependencies for XML, in: Proc. EDBT, 2002, pp. 124–141.
- [38] S. Hartmann, S. Link, More functional dependencies for xml, in: L. Kalinichenko, R. Manthey, B. Thalheim, U. Wloka (Eds.), Advances in Databases and Information Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 355–369.
- [39] L. Kot, W. M. White, Characterization of the interaction of XML functional dependencies with DTDs, in: Proc. ICDT, 2007, pp. 119–133.
- [40] S. Kruse, T. Papenbrock, F. Naumann, Scaling out the discovery of inclusion dependencies, in: Proc. BTW, 2015, pp. 445–454.
- [41] T. Vajk, P. Fehér, K. Fekete, H. Charaf, Denormalizing data into schema-free databases, in: Proc. CogInfoCom, 2013, pp. 747–752.
- [42] X. Chu, I. F. Ilyas, S. Krishnan, J. Wang, Data cleaning: Overview and emerging challenges, in: Proc. SIGMOD, 2016, pp. 2201–2206.
- [43] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, A. Grafberger, Automating large-scale data quality verification, Proc. VLDB Endow. 11 (2018) 1781–1794.
- [44] J. Kivinen, H. Mannila, Approximate inference of functional dependencies from relations, Theoretical Computer Science 149 (1995) 129–149.
- [45] H. Mannila, K.-J. Räihä, On the complexity of inferring functional dependencies, Discrete Applied Mathematics 40 (1992) 237–243.
- [46] H. Yao, H. J. Hamilton, Mining functional dependencies from data, Data Mining and Knowledge Discovery 16 (2008) 197–219.
- [47] T. Papenbrock, F. Naumann, A hybrid approach to functional dependency discovery, in: Proc. SIGMOD, 2016, pp. 821–833.
- [48] J. Bauckmann, U. Leser, F. Naumann, V. Tietz, Efficiently detecting inclusion dependencies, in: Proc. ICDE, 2007, pp. 1448–1450.
- [49] R. Hai, C. Quix, D. Wang, Relaxed functional dependency discovery in heterogeneous data lakes, in: Proc. ER, 2019, pp. 225–239.
- [50] W. Fan, F. Geerts, L. V. S. Lakshmanan, M. Xiong, Discovering conditional functional dependencies, in: Proc. ICDE, 2009, pp. 1231–1234.
- [51] W. Fan, F. Geerts, J. Li, M. Xiong, Discovering conditional functional dependencies, IEEE Trans. Knowl. Data Eng. 23 (2011) 683–698.

- [52] J. Li, J. Liu, H. Toivonen, J. Yong, Effective pruning for the discovery of conditional functional dependencies, *Comput. J.* 56 (2013) 378–392.
- [53] I. F. Ilyas, X. Chu, *Data Cleaning*, ACM, 2019.
- [54] P. Schirmer, T. Papenbrock, S. Kruse, F. Naumann, D. Hempfing, T. Mayer, D. Neuschäfer-Rube, Dynfd: Functional dependency discovery in dynamic datasets, in: *Proc. EDBT*, 2019, pp. 253–264.
- [55] Z. Wei, S. Link, Discovery and ranking of functional dependencies, in: *Proc. ICDE*, 2019, pp. 1526–1537.
- [56] A. A. Frozza, R. dos Santos Mello, F. de Souza da Costa, An approach for schema extraction of JSON and extended JSON document collections (software), <https://github.com/gbd-ufsc/JSONSchemaDiscovery>, 2018.
- [57] J. Liu, J. Li, C. Liu, Y. Chen, Discover dependencies from data - A review, *IEEE Trans. Knowl. Data Eng.* 24 (2012) 251–264.