F# Data: Making structured data first-class citizens

Tomas Petricek
University of Cambridge
tomas@tomasp.net

Gustavo Guerra
Microsoft Corporation, London
gustavo@codebeside.org

Don Syme
Microsoft Research, Cambridge
dsyme@microsoft.com

Abstract

Most statically typed languages assume that programs run in a closed world, but this is not the case. Modern applications interact with external services and often access data in structured formats such as XML, CSV and JSON. Static type systems do not understand such external data sources and only make data access more cumbersome. Should we just give up and leave the messy world of external data to dynamic typing and runtime checks?

Of course, we should not give up! In this paper, we show how to integrate external data sources into the F# type system. As most real-world data on the web do not come with an explicit schema, we develop a type inference algorithm that infers a type from representative samples. Next, we use the type provider mechanism for integrating the inferred structures into the F# type system.

The resulting library significantly reduces the amount of code that developers need to write when accessing data. It also provides additional safety guarantees. Arguably, as much as possible if we abandon the incorrect closed world assumption.

Categories and Subject Descriptors D.3.3 [*Programming Languages*]: Language Constructs and Features

Keywords F#, Type Providers, JSON, XML, Data, Type Inference

1. Introduction

The key function of many modern applications is to connect multiple data sources or services and present the aggregated data in some form. Mobile applications for taking notes, searching train schedules or finding tomorrow's weather all communicate with one or more services over the internet.

Increasing number of such services provide REST-based endpoints that return data as CSV, XML or JSON. Despite numerous schematization efforts, the services typically do not come with schema. At best, the documentation provides a number of typical requests and sample responses.

For example, the OpenWeatherMap service provides an endpoint for getting current weather for a given city[1]. The page documents the input URL parameters and shows one sample JSON response to illustrate the response structure. Using a standard library for working with JSON and HTTP, we might call the service and read the temperature as follows [2]:

```
let data = Http.Request("http://weather.org/?q=Prague")
match JsonValue.Parse(data) with
| Record(root) →
  match Map.find "main" root with
  | Record(main) →
    match Map.find "temp" main with
    | Number(num) → printfn "Lovely %f degrees!" num
    | _ → failwith "Incorrect format"
  | _ → failwith "Incorrect format"
| _ → failwith "Incorrect format"
```

The code assumes that the response has a particular format described in the documentation. The root node must be a record with a "main" field, which has to be another record containing a "temp" field with numerical value. When the format is incorrect, the data access simply fails with an exception.

Using the JSON type provider from the F# Data library, we can write code with exactly the same functionality in two lines:

```
type W = JsonProvider⟨"http://weather.org/?q=Prague"⟩
printfn "Lovely %f degrees!" (W.GetSample().Main.Temp)
```

On the first line, JsonProvider⟨"..."⟩ invokes a type provider at compile-time with the URL as a sample. The type provider infers the structure of the response and provides a type with a GetSample method that returns a parsed JSON with nested properties Main.Temp, returning the temperature as a number.

In the rest of the paper, we give more detailed description of the mechanism and we also discuss theoretical safety properties of the approach. The key novel contributions of this paper are:

- We present type providers for XML, CSV and JSON (Section 2) that are available in the F# Data library and we also cover practical aspects of the implementation that contributed to their industrial adoption (Section 6).
- We describe a predictable type inference algorithm for structured data formats that underlies the type providers (Section 3). It is based on finding common supertype of a set of examples.
- We provide a minimal formal model (Section 4) and use it to prove a relativized notion of type safety for our type providers (Section 5). Although we focus on our specific case, the relativized safety property demonstrates a new way of thinking about ML-style type safety that is needed in the context of web.

Although the F# Data library [16] has been widely adopted is one of the most downloaded F# libraries it has not yet been presented in an academic form. Additional documentation and source code can be found at http://fsharp.github.io/FSharp.Data.

[1] See "Current weather data": http://openweathermap.org/current
[2] We abbreviate the full URL: http://api.openweathermap.org/data/2.5/weather?q=Prague&units=metric

2. Structural type providers

We start with an informal overview that shows how F# Data type providers simplify working with JSON, XML and CSV. We also introduce the necessary aspects of F# 3.0 type providers along the way. The examples in this section illustrate a number of key properties of our type inference algorithm:

- Our mechanism is robust and predictable. This is important as the user directly works with the inferred types and should understand why a specific type was inferred from a given sample[3].
- Our inference mechanism prefers records over unions. This better supports developer tooling – most F# editors provide code completion hints on "." and so types with properties (records) are easier to use than types that require pattern matching.
- Finally, we handle a number of practical concerns that may appear overly complicated, but are important in the real world. This includes support for different numerical types, null values and missing data and also different ways of representing Booleans in CSV files.

2.1 Working with JSON documents

The JSON format used in the example in Section 1 is a popular format for data exchange on the web based on data structures used in JavaScript. The following is the definition of the JsonValue type used earlier to represent parsed JSON:

```
type JsonValue =
  | Null
  | Number of decimal
  | String of string
  | Boolean of bool
  | Record of Map⟨string, JsonValue⟩
  | Array of JsonValue[]
```

The OpenWeatherMap example in the introduction used only a (nested) record containing a numerical value. To demonstrate other aspects of the JSON type provider, we look at a more complex example that also involves null value and an array:

```
[ { "name": null, "age": 25 },
  { "name": "Alexander", "age": 3.5 },
  { "name": "Tomas" } ]
```

Say we want to print the names of people in the list with an age if it is available. As before, the standard approach would be to pattern match on the parsed JsonValue. The code would check that the top-level node is a Array, iterate over the elements checking that each is a Record with certain properties and throw an exception or skip values in incorrect format. The standard approach can be made nicer by defining helper functions. However, we still need to specify names of fields as strings, which is error prone and can not be statically checked.

Assuming the people.json file contains the above example and data is a string value that contains another data set in the same format, we can print names and ages as follows:

```
type People = JsonProvider⟨"people.json"⟩

let items = People.Parse(data)
for item in items do
  printf "%s " item.Name
  Option.iter (printf "(%f)") item.Age
```

The code achieves the same simplicity as when using dynamically typed languages, but is statically type-checked. In contrast to the earlier example, we now use a local file people.json as a sample for the type inference, but then processes data from another source.

Type providers. The notation JsonProvider⟨"people.json"⟩ on the first line passes a *static parameter* to the type provider. Static parameters are resolved at compile-time, so the file name has to be a constant. The provider analyzes the sample and generates a type that we name People. In F# editors that use the F# Compiler Service, the type provider is also executed at development-time and so the same provided types are used in code completion.

The JsonProvider uses a type inference algorithm (Section 3) and infers the following types from the sample:

```
type Entity =
  member Age : option⟨decimal⟩
  member Name : option⟨string⟩

type People =
  member GetSample : unit → Entity[]
  member Parse : string → Entity[]
```

The type Entity represents the person. The property Name is optional, because one of the sample records contains null as the name. The property Age is also optional, because the value is missing in one sample. The two sample ages are an integer 25 and a decimal 3.5 and so the common inferred type is decimal.

The type People provides two methods for reading data. GetSample returns the sample used for the inference and Parse parses a JSON string containing data in the same format as the sample. Since the sample JSON is a collection of records, both methods return an array of Entity values.

Error handling. In addition to the structure of the types, the type provider also specifies what code should be executed at run-time in place of item.Name and other operations. This is done through a *type erasure* mechanism demonstrated in Section 2.2. In this example, the runtime behaviour is the same as in the hand-written sample in Section 1 – a member access throws an exception if the document does not have the expected format.

Informally, the safety property discussed in Section 5 states that if the inputs are subtypes of some of the provided samples (*i.e.* the samples are representative), then no exceptions will occur. In other words, we cannot avoid all failures, but we can prevent some – for example, if the OpenWeatherMap changes the response format, the sample in Section 1 will not re-compile and the user knows that the code needs to be changed. What this means for the traditional ML type safety is discussed in Section 5.1.

The role of records. The sample code is easy to write thanks to the fact that most F# editors provide code completion when "." is typed. The developer does not need to look at the sample JSON file to see what fields are available for a person. To support this scenario, our inference algorithm prefers records (we also treat XML elements and CSV rows as records).

In the above example, this is demonstrated by the fact that Age is marked as optional. An alternative is to provide two different record types (one with Name and other with Name and Age), but this would complicate the processing code.

2.2 Reading CSV files

In the CSV file format, the structure is a collection of rows (records) consisting of fields (with names specified by the first row). The inference needs to infer the types of fields. For example:

```
Ozone, Temp, Date,       Autofilled
41,    67,   2012-05-01, 0
36.3,  72,   2012-05-02, 1
12.1,  74,   3 kveten,   0
17.5,  #N/A, 2012-05-04, 0
```

[3] In particular, we do not use probabilistic methods where adding additional sample could change the shape of the type.

One difference between JSON and CSV formats is that in CSV, the literals have no data types. In JSON, strings are "quoted" and Booleans are true and false. This is not the case for CSV and so we need to infer not just the structure, but also the primitive types.

Assuming the sample is saved as airdata.csv, the following snippet prints all rows from another file that were not autofilled:

```
type AirCsv = CsvProvider⟨"airdata.csv"⟩
let air = AirCsv.Parse(data)
for row in air.Rows do
  if not row.Autofilled then
    printf "%s: %d" row.Date row.Ozone
```

The type of the record (row) is, again, inferred from the sample. The Date column uses mixed formats and is inferred as string (although we support many date formats and "May 3" would work). More interestingly, we also infer Autofilled as Boolean, because the sample contains only 0 and 1 values and using those for Booleans is a common CSV convention. Finally, the fields Ozone and Temp have types decimal and option⟨int⟩.

Erasing type providers. At runtime, the type providers we describe use an erasure mechanism similar to Java Generics [1]. A type provider also generates code that is executed in place of the members of the generated types. In the above example, the compiled (and actually executed) code looks as follows:

```
let air = CsvFile.Load("airdata.csv", fun r →
  asDecimal r.[0], asIntOpt r.[1], r.[2], asBool r.[3])
for c1, _, c3, c4 in air.Rows do
  if not c4 then printf "%s: %d" c3 c1
```

The generated type AirCsv is erased to a type CsvFile⟨α⟩. The Load method takes the file name together with a function that turns a row represented as an array of strings to a typed representation – here, a four-element tuple decimal∗option⟨int⟩∗string∗bool (this type is also used as a type argument of CsvFile⟨α⟩. The properties such as Ozone are then replaced with an accessor that reads the corresponding tuple element.

The type erasure mechanism is explained in our simplified formal model in Section 4. More details about the full implementation in F# can be found in the paper on F# type providers [23].

2.3 Processing XML data

The XML format is similar to JSON in that it uses nested elements rather than a flat CSV-like structure. In our inference algorithm, we use the notion of *records* for both JSON records and XML elements. However, unlike in JSON, the XML elements also have a name. An element can contain a collection of child elements. In case of XML, the collection can often be heterogeneous containing child nodes of different names. Consider the following (simplified) RSS news feed as an example:

```
<rss version="2.0"><channel>
  <title> BBC News - Europe </title>
  <item><title> Kurdish activists
      killed in Paris </title></item>
  <item><title> German MPs warn
      over UK EU exit </title></item>
</channel></rss>
```

The channel node contains a single title node (with the name of the feed) and multiple item nodes (individual news articles). In RSS, the title nodes always contain plain text, while item nodes contain title (and also url and other properties omitted here).

When extracting data from such XML documents, we often need to get nodes of a specific name (all items) or extract the content of a node (feed title). The XML type provider is optimized for this style of access and allows processing the feed as follows:

```
type RssFeed = XmlProvider⟨"rss-sample.xml"⟩
let channel = RssFeed.Load("http://bbc.co.uk/...").Channel
printf "News from: %s " channel.Title
for item in channel.Items do
  printf " - %s " item.Title
```

In this example, the type of the channel value represents a heterogeneous collection that contains exactly one title node and zero or more item nodes.

Heterogeneous collections and unions. Our inference algorithms recognizes this as a *heterogeneous collection* (if the same structure appears in all samples) and generates a type with a member Title for accessing the value of the singleton title element together with a member Items that returns a collection of zero or more items. As a further simplification, if an element contains only a primitive value (such as title) it is represented as a value of primitive type (so channel.Title has a type string).

As an alternative, the inference could produce a collection of union type (with a case for item and a case for title). This alternative is useful for documents with complicated structure (such as XHTML files), but less useful for files with simpler structure such as web server responses, data stores and configuration files.

Moreover, our approach based on heterogeneous collections has a nice property that it generates types with the same public interface for the following two ways of representing data in XML:

```
<author name="Tomas" age="27" />
<author><name>Tomas</name><age>27</age></author>
```

In both cases, the provided type for author has two properties, typed string and int respectively. In Section 3, we discuss the simplified model (using union types). The inference based on heterogeneous collections is described later in Section 6.2.

2.4 Summary

The previous three sections demonstrated the F# Data type providers for JSON, CSV and XML. A reader interested in more examples is invited to look at documentation on the F# Data web site[4].

It should be noted that we did not attempt to present the library using well-structured ideal samples, but instead used data sets that demonstrate typical issues that are frequent in real world inputs (missing data, inconsistent encoding of Booleans and heterogeneous structures).

From looking at just the previous three examples, these may appear arbitrary, but our experience suggests that they are the most common issues. The following JSON response with government debt information returned by the World Bank API demonstrates all three problems:

```
[ { "page": 1, "pages": 5 },
  [ { "indicator": "GC.DOD.TOTL.GD.ZS",
      "date": "2012",
      "value": null },
    { "indicator": "GC.DOD.TOTL.GD.ZS",
      "date": "2010",
      "value": "35.1422970266502" } ] ]
```

First of all, the top-level element is a collection containing two values of different kind. The first is a record with meta-data about the current page and the second is an array with data. The JSON type provider infers this as heterogeneous collection with properties Record for the former and Array for the latter. Second, the value is null for some records. Third, numbers can be represented in JSON as numeric literals (without quotes), but here, they are returned as string literals instead[5].

[4] Available at http://fsharp.github.io/FSharp.Data/

[5] This is often used to avoid non-standard numerical types in JavaScript.

3. Structural type inference

Our type inference algorithm for structured formats is based on a subtyping relation. When inferring the type of a document, we infer the most specific types of individual values (CSV rows, JSON or XML nodes) and find a common supertype of values in the sample.

In this section, we define the *structural type* σ, which is a type of structured data. Note that this type is distinct from the programming language types τ (the type generates the latter from the former). Next, we define the subtyping relation on structural types σ and describe the algorithm for finding a common supertype.

Note that the subtyping relation between structural types σ does not map to a subtyping relation between F# types τ. However, it does hold at runtime, meaning that an operation on a supertype can be performed on its subtypes. We make the type provider mechanism and runtime aspects precise in Section 4 before discussing safety in Section 5.

3.1 Structural types

The grammar below defines *structural type* σ. We distinguish between *non-nullable types* that always have a valid value (written as $\hat{\sigma}$) and *nullable types* that encompass missing and null values (written as σ). We write ν for record field names and $\nu?$ for record names, which can be empty:

$$\hat{\sigma} = \nu?\ \{\nu_1 : \sigma_1, \ldots, \nu_n : \sigma_n\} \mid \mathsf{float} \mid \mathsf{decimal} \mid \mathsf{int} \mid \mathsf{bit} \mid \mathsf{bool} \mid \mathsf{string}$$

$$\sigma = \hat{\sigma} \mid \hat{\sigma}\ \mathsf{option} \mid [\sigma] \mid \sigma_1 + \ldots + \sigma_n \mid \top \mid \mathsf{null}$$

The non-nullable types include records (consisting of an optional name and zero or more fields with their types) and primitive types. Records arising from XML documents are always named, while records used by JSON and CSV providers are always unnamed.

Next, we have three numerical types – int for integers, decimal for small high-precision decimal numbers and float for floating-point numbers (these are related by sub-typing relation as discussed in the next section). Furthermore, the bit type is a type of values 0 and 1 and makes it possible to infer Boolean in the CSV processing example discussed in Section 2.2.

Any non-nullable type is also a nullable type, but it can be wrapped in the option constructor to explicitly permit the null value. These are typically mapped to the standard F# option type. A simple collection type $[\sigma]$ is also nullable and missing values or null are treated as empty collection. The type null is inhabited by the null value (using an overloaded but not ambiguous notation) and $\top$ represents the top type.

Finally, a union type in our model implicitly permits the null value. This is because structural union types are *not* mapped to F# union types. Instead our encoding requires the user to handle the cases one-by-one and so they also need to handle the situation when none of the cases matches (as discussed in Section 6.3, this is the right choice for an F# type provider, but it would not necessarily be a good fit for other languages).

3.2 Subtyping relation

To provide a basic intuition, the subtyping relation between structural types is illustrated in Figure 1. We split the diagram in two parts. The upper part shows non-nullable types (with records and primitive types). The lower part shows nullable types with null, collections and optional values. We omit links between the two part, but any type $\hat{\sigma}$ is a subtype of $\hat{\sigma}$ option (in the diagram, we abbreviate σ option as $\sigma?$).

The diagram shows only the basic structure and it does not explain relationships between records with the same name and between union types. This following definition makes this precise.

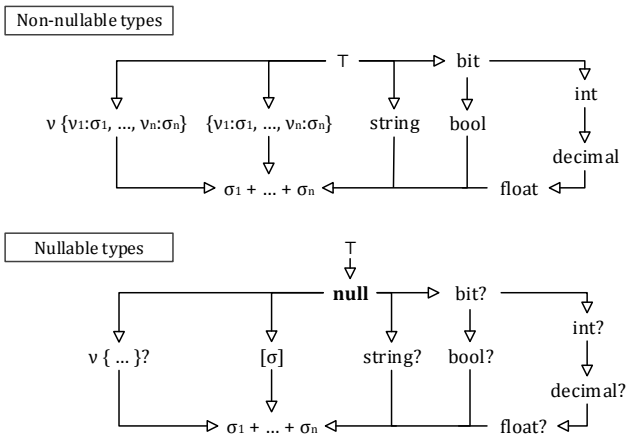

Figure 1. Subtype relation between structural types

Definition 1. *We write $\sigma_1 :> \sigma_2$ to denote that σ_2 is a subtype of σ_1. The subtyping relation is defined as a transitive reflexive closure of these rules:*

Primitives, options, collections

$$\mathsf{float} :> \mathsf{decimal} :> \mathsf{int} :> \mathsf{bit} \quad \text{and} \quad \mathsf{bool} :> \mathsf{bit} \qquad \text{(B1)}$$
$$\sigma :> \mathsf{null} \quad (\text{iff } \sigma \neq \hat{\sigma}) \qquad \text{(B2)}$$
$$\hat{\sigma}\ \mathsf{option} :> \hat{\sigma} \quad (\text{for all } \hat{\sigma}) \qquad \text{(B3)}$$
$$\sigma :> \top \quad (\text{for all } \sigma) \qquad \text{(B4)}$$
$$[\sigma_1] :> [\sigma_2] \quad (\text{if } \sigma_1 :> \sigma_2) \qquad \text{(B5)}$$

Union types

$$\sigma_1 + \ldots + \sigma_n :> \sigma'_1 + \ldots + \sigma'_n \quad (\text{iff } \forall i.\sigma_i :> \sigma'_i) \qquad \text{(U1)}$$
$$\sigma_1 + \ldots + \sigma_n :> \sigma_1 + \ldots + \sigma_m \quad (\text{where } m \leq n) \qquad \text{(U2)}$$
$$\sigma_1 + \ldots + \sigma_n :> \sigma_1 + \ldots + \sigma_m \quad (\text{where } m \geq n) \qquad \text{(U3)}$$
$$\sigma_1 + \ldots + \sigma_n :> \sigma_{\pi(1)} + \ldots + \sigma_{\pi(n)} \quad (\text{permutation } \pi) \qquad \text{(U4)}$$
$$\sigma_1 + \ldots + \sigma_n :> \sigma_i \quad (\text{for } i \in 1 \ldots n) \qquad \text{(U5)}$$

Record types

$$\nu?\ \{\nu_1 : \sigma_1, .., \nu_n : \sigma_n\} :> \nu?\ \{\nu_1 : \sigma'_1, .., \nu_n : \sigma'_n\}\ (\sigma_i :> \sigma'_i) \qquad \text{(R1)}$$

$$\nu?\ \{\nu_1 : \sigma_1, .., \nu_n : \sigma_n\} :> \nu?\ \{\nu_1 : \sigma_1, .., \nu_m : \sigma_m\}\ (m \geq n) \qquad \text{(R2)}$$

$$\nu?\ \{\nu_1 : \sigma_1, .., \nu_n : \sigma_n\} :> \nu?\ \{\nu_{\pi(1)} : \sigma_{\pi(1)}, .., \nu_{\pi(m)} : \sigma_{\pi(m)}\} \quad (\text{permutation } \pi) \qquad \text{(R3)}$$

$$\nu?\ \{\nu_1 : \sigma_1, .., \nu_n : \sigma_n, \nu_{n+1} : \mathsf{null}, .., \nu_{n+m} : \mathsf{null}\} :> \nu?\ \{\nu_1 : \sigma_1, .., \nu_n : \sigma_n\} \qquad \text{(R4)}$$

Here is a summary of the key aspects of the definition:

- For numeric types (B1), we infer a single most precise numeric type that can represent all values from a sample dataset, so we order types by the ranges they represent; int is a 32-bit integer, decimal has a range of about -1^{29} to 1^{29} and float is a double-precision floating-point.
- The bit type is a subtype of both int and bool (B1). Thus, a sample $0, 1$ has a type bit; $0, 1,$ true is bool and $0, 1, 2$ becomes int (for a sample $0, 1, 2,$ true we would need a union type).
- The null type is a subtype of all nullable types (B2), that is all σ types excluding non-nullable types $\hat{\sigma}$). Any non-nullable type is also a subtype of its optional version (B3).

$$\text{(record-1)}\ \frac{(\nu_i = \nu'_j \Leftrightarrow (i = j) \wedge (i \leq k)) \qquad \forall i \in \{1..k\}.(\sigma_i \bigtriangledown \sigma'_i \vdash \sigma''_i)}{\nu?\ \{\nu_1 : \sigma_1,\ \ldots\ , \nu_k : \sigma_k,\ \ldots\ , \nu_n : \tau_n\} \bigtriangledown \nu?\ \{\nu'_1 : \sigma'_1,\ \ldots\ , \nu'_k : \sigma'_k,\ \ldots\ , \nu'_m : \tau'_m\} \vdash \nu?\ \{\nu_1 : \sigma''_1,\ \ldots\ , \nu_k : \sigma''_k, \nu_{k+1} : \lceil\sigma_{k+1}\rceil, \ldots, \nu_n : \lceil\sigma_n\rceil, \nu'_{k+1} : \lceil\sigma'_{k+1}\rceil, \ldots, \nu'_m : \lceil\sigma'_m\rceil\}}$$

$$\text{(union-1)}\ \frac{\exists i.\mathsf{tagof}(\sigma_i) = \mathsf{tagof}(\sigma) \quad \sigma \bigtriangledown \sigma_i \vdash \sigma'_i \quad \mathsf{tagof}(\sigma) \neq \mathsf{union}}{\sigma \bigtriangledown (\sigma_1 + \ldots + \sigma_n) \vdash (\sigma_1 + \ldots + \lfloor\sigma'_i\rfloor + \ldots + \sigma_n)} \qquad \frac{\nexists i.\mathsf{tagof}(\sigma_i) = \mathsf{tagof}(\sigma) \quad \mathsf{tagof}(\sigma) \neq \mathsf{union}}{\sigma \bigtriangledown (\sigma_1 + \ldots + \sigma_n) \vdash (\sigma_1 + \ldots + \sigma_n + \lfloor\sigma\rfloor)}$$

$$\text{(union-2)}\ \frac{\mathsf{tagof}(\sigma_i) = \mathsf{tagof}(\sigma'_i) \quad \sigma_i \bigtriangledown \sigma'_i \vdash \sigma''_i \qquad (\forall i \in \{1..k\})}{(\sigma_1 + \ldots + \sigma_k + \ldots + \sigma_n) \bigtriangledown (\sigma'_1 + \ldots + \sigma'_k + \ldots + \sigma'_m) \vdash (\sigma''_1 + \ldots + \sigma''_k + \sigma_{k+1} + \ldots + \sigma_n + \sigma'_{k+1} + \ldots + \sigma'_m)} \qquad \text{(union-3)}\ \frac{\text{(no other rule applies)}}{\sigma_1 \bigtriangledown \sigma_2 \vdash \lfloor\sigma_1\rfloor + \lfloor\sigma_2\rfloor}$$

$$\text{(order-1)}\ \frac{\nu?\ \{\nu_1 : \sigma_1, \ldots, \nu_n : \sigma_n\} \bigtriangledown \sigma \vdash \sigma'}{\nu?\ \{\nu_{\pi(1)} : \sigma_{\pi(1)}, \ldots, \nu_{\pi(n)} : \sigma_{\pi(n)}\} \bigtriangledown \sigma \vdash \sigma'} \qquad \text{(order-2)}\ \frac{\sigma_1 + \ldots + \sigma_n \bigtriangledown \sigma \vdash \sigma'}{\sigma_{\pi(1)} + \ldots + \sigma_{\pi(n)} \bigtriangledown \sigma \vdash \sigma'} \qquad (\pi \text{ permutation})$$

$$\text{(list)}\ \frac{\sigma_1 \bigtriangledown \sigma_2 \vdash \sigma}{[\sigma_1] \bigtriangledown [\sigma_2] \vdash [\sigma]} \qquad \text{(prim)}\ \frac{\sigma_1 :> \sigma_2}{\sigma_1 \bigtriangledown \sigma_2 \vdash \sigma_1} \quad \begin{array}{l}(\text{when } \sigma_1, \sigma_2 \in \{\mathsf{bit}, \mathsf{int}, \mathsf{decimal}, \mathsf{float}\} \\ \quad \text{or } \sigma_1, \sigma_2 \in \{\mathsf{bit}, \mathsf{bool}\})\end{array}$$

$$\text{(sym)}\ \frac{\sigma_1 \bigtriangledown \sigma_2 \vdash \sigma}{\sigma_2 \bigtriangledown \sigma_1 \vdash \sigma} \qquad \text{(refl)}\ \sigma \bigtriangledown \sigma \vdash \sigma \qquad \text{(null-1)}\ \sigma \bigtriangledown \mathsf{null} \vdash \sigma \quad (\sigma :> \mathsf{null})$$

$$\text{(top)}\ \top \bigtriangledown \sigma \vdash \sigma \qquad \text{(null-2)}\ \sigma \bigtriangledown \mathsf{null} \vdash \sigma\ \mathsf{option} \quad (\sigma :\not> \mathsf{null})$$

Figure 2. Inference judgements that define the common supertype relation

- There is a top type (B4), but no bottom type. It is possible to find common supertype of any two types, because a union type $\tau_1 + \ldots + \tau_n$ is a supertype of all its components (U5).
- For unions, we include the standard rules. Subtype can have fewer alternatives (U1), elements can be reordered (U2) and subtyping is covariant (U4). An unusual rule (U3) says that subtype can also have *more* cases. This is valid because the type provider exposes unions as objects that do not allow complete pattern matching.
- As usual, the subtyping on records is covariant (R1), subtype can have additional fields (R2) and fields can be reordered (R3). The interesting rule is the last one (R4). Together with covariance, it states that a subtype can omit some fields, provided that their types are nullable.

The rule that allows subtype to have fewer record elements (R4) is particularly important. It allows us to prefer records in some cases. For example, given two samples $\{\mathsf{name} : \mathsf{string}\}$ and $\{\mathsf{name} : \mathsf{string}, \mathsf{age} : \mathsf{int}\}$, we can find a common supertype $\{\mathsf{name} : \mathsf{string}, \mathsf{age} : \mathsf{int}\ \mathsf{option}\}$ which is also a record. For usability reasons, we prefer this to another common supertype $\{\mathsf{name} : \mathsf{string}\} + \{\mathsf{name} : \mathsf{string}, \mathsf{age} : \mathsf{int}\}$. The next section describes precisely how our inference algorithm works.

It is worth noting that some of the aspects (such as 3 different numeric types and handling of the null type) are specific to F#. Similar problems will appear with most real-world languages and systems. JVM and OCaml all haves multiple numeric types, but they would handle null differently.

We could also be more or less strict about missing data – we choose to handle missing values silently when we can (null becomes an empty collection), but we are explicit in other cases (we infer string option as a hint to the user rather than treating null as an empty string). These choices are based on experience with using F# Data in practice.

3.3 Common supertype relation

As mentioned earlier, the structural type inference relies on a finding common supertype. However, the partially ordered set of types does not have a unique greatest lower bound. For example, record types $\{\mathsf{a} : \mathsf{int}\}$ and $\{\mathsf{b} : \mathsf{bool}\}$ have common supertypes $\{?\mathsf{a} : \mathsf{int}, ?\mathsf{b} : \mathsf{bool}\}$ and $\{\mathsf{a} : \mathsf{int}\} + \{\mathsf{b} : \mathsf{bool}\}$ which are unrelated. Our inference algorithm always prefers records over unions (Theorem 3). The definition of the *common supertype* relation uses a number of auxiliary definitions that are explained in the discussion that follows.

Definition 2. *A* common supertype *of types σ_1 and σ_2 is a type σ, written $\sigma_1 \bigtriangledown \sigma_2 \vdash \sigma$, obtained according to the inference rules in Figure 2.*

When finding a common supertype of two records (*record-1*), we return a record type that has the union of fields of the two arguments. We assume that the names of the first k fields are the same and the remaining fields have different names. To get a record with the right order of elements, we provide the (*order*) rule. The types of shared fields become common supertypes of their respective types (recursively). Fields that are present in only one record are marked as optional using the following helper definition:

$$\begin{array}{ll}\lceil\hat{\sigma}\rceil = \hat{\sigma}\ \mathsf{option} & \text{(non-nullable types)} \\ \lceil\sigma\rceil = \sigma & \text{(otherwise)}\end{array}$$

Finding a common supertype of when one type is a union is more complicated. Our definition aims to use a flat structure (avoid nesting unions) and limit the number of cases. This is done by grouping types that have a common supertype which is not a union type. For example, rather than inferring $\mathsf{int} + (\mathsf{bool} + \mathsf{decimal})$, our algorithm will find the common supertype of int and decimal (which is decimal) and it will produce just $\mathsf{decimal} + \mathsf{bool}$.

To identify which types have a common supertype which is not a union, we group the types by a *tag*, which is defined as:

$$\begin{array}{rl}\mathsf{tag} = & \mathsf{string} \mid \mathsf{bool} \mid \mathsf{number} \mid \mathsf{union} \mid \mathsf{collection} \\ \mid & \mathsf{rec\text{-}anon} \mid \mathsf{rec\text{-}named}\ \nu\end{array}$$

The tag of a type is obtained using a function $\mathsf{tagof}(-) : \sigma \to \mathsf{tag}$:

$$\begin{array}{rl}\mathsf{tagof}(\mathsf{string}) &= \mathsf{string} \\ \mathsf{tagof}(\mathsf{bool}) &= \mathsf{bool} \\ \mathsf{tagof}(\mathsf{decimal}) = \mathsf{tagof}(\mathsf{float}) &= \mathsf{number} \\ \mathsf{tagof}(\mathsf{int}) = \mathsf{tagof}(\mathsf{bit}) &= \mathsf{number} \\ \mathsf{tagof}([\sigma]) &= \mathsf{collection}\end{array}$$

$\mathsf{asInt}(i) \rightsquigarrow i$
$\mathsf{asInt}(d) \rightsquigarrow i \quad (i = \lfloor d \rfloor)$
$\mathsf{asInt}(f) \rightsquigarrow i \quad (i = \lfloor f \rfloor)$
$\mathsf{asInt}(t) \rightsquigarrow i \quad (t \text{ represents } i)$
$\mathsf{asInt}(\mathsf{true}) \rightsquigarrow 1$
$\mathsf{asInt}(\mathsf{false}) \rightsquigarrow 0$

$\mathsf{asDec}(i) \rightsquigarrow d \quad (d = i)$
$\mathsf{asDec}(d) \rightsquigarrow d$
$\mathsf{asDec}(f) \rightsquigarrow d \quad (d = \lfloor f \rfloor)$
$\mathsf{asDec}(t) \rightsquigarrow d \quad (t \text{ represents } d)$
$\mathsf{asDec}(\mathsf{true}) \rightsquigarrow 1.0$
$\mathsf{asDec}(\mathsf{false}) \rightsquigarrow 0.0$

$\mathsf{asFloat}(i) \rightsquigarrow f \quad (f = i)$
$\mathsf{asFloat}(d) \rightsquigarrow f \quad (f = d)$
$\mathsf{asFloat}(f) \rightsquigarrow f$
$\mathsf{asFloat}(t) \rightsquigarrow f \quad (t \text{ represents } f)$
$\mathsf{asFloat}(\mathsf{true}) \rightsquigarrow 1.0$
$\mathsf{asFloat}(\mathsf{false}) \rightsquigarrow 0.0$

$\mathsf{asStr}(i) \rightsquigarrow t \quad (t \text{ represents } i)$
$\mathsf{asStr}(d) \rightsquigarrow t \quad (t \text{ represents } d)$
$\mathsf{asStr}(f) \rightsquigarrow t \quad (t \text{ represents } f)$
$\mathsf{asStr}(t) \rightsquigarrow t$
$\mathsf{asStr}(\mathsf{true}) \rightsquigarrow \text{"true"}$
$\mathsf{asStr}(\mathsf{false}) \rightsquigarrow \text{"false"}$

$\mathsf{asBool}(1) \rightsquigarrow \mathsf{true}$
$\mathsf{asBool}(1.0) \rightsquigarrow \mathsf{true}$
$\mathsf{asBool}(\text{"true"}) \rightsquigarrow \mathsf{true}$
$\mathsf{asBool}(\text{"1"}) \rightsquigarrow \mathsf{true}$

$\mathsf{asBool}(0) \rightsquigarrow \mathsf{false}$
$\mathsf{asBool}(0.0) \rightsquigarrow \mathsf{false}$
$\mathsf{asBool}(\text{"false"}) \rightsquigarrow \mathsf{false}$
$\mathsf{asBool}(\text{"0"}) \rightsquigarrow \mathsf{false}$
$\mathsf{asBool}(b) \rightsquigarrow b$

$\mathsf{isNull}(\mathsf{null}) \rightsquigarrow \mathsf{true}$
$\mathsf{isNull}(\_) \rightsquigarrow \mathsf{false}$

$\mathsf{getChildren}(\mathsf{null}) \rightsquigarrow [\,]$

$\mathsf{getField}(\nu, \nu_i, \nu\ \{\ldots, \nu_i = s_i, \ldots\}) \rightsquigarrow s_i$
$\mathsf{getField}(\bot, \nu_i, \{\ldots, \nu_i = s_i, \ldots\}) \rightsquigarrow s_i$
$\mathsf{getField}(\nu, \nu', \nu\ \{\ldots, \nu_i = s_i, \ldots\}) \rightsquigarrow \mathsf{null} \quad (\nexists i.\nu_i = \nu')$
$\mathsf{getField}(\bot, \nu', \{\ldots, \nu_i = s_i, \ldots\}) \rightsquigarrow \mathsf{null} \quad (\nexists i.\nu_i = \nu')$

$\mathsf{getChildren}([s_1; \ldots; s_n]) \rightsquigarrow [s_1; \ldots; s_n]$

$\mathsf{isTag}(\mathsf{string}, t) \rightsquigarrow \mathsf{true}$
$\mathsf{isTag}(\mathsf{bool}, v) \rightsquigarrow \mathsf{true} \quad (\text{when } v \in 0, 1, 0.0, 1.0, \mathsf{true}, \mathsf{false})$
$\mathsf{isTag}(\mathsf{number}, v) \rightsquigarrow \mathsf{true} \quad (\text{when } v = i, v = d, v = f)$
$\mathsf{isTag}(\mathsf{collection}, [s_1; \ldots; s_n]) \rightsquigarrow \mathsf{true}$

$\mathsf{isTag}(\mathsf{collection}, \mathsf{null}) \rightsquigarrow \mathsf{true}$
$\mathsf{isTag}(\text{rec-anon}, \{\nu_1 \mapsto s_1, \ldots, \nu_n \mapsto s_n\}) \rightsquigarrow \mathsf{true}$
$\mathsf{isTag}(\text{rec-named } \nu, \nu\ \{\nu_1 \mapsto s_1, \ldots, \nu_n \mapsto s_n\}) \rightsquigarrow \mathsf{true}$
$\mathsf{isTag}(\_, \_) \rightsquigarrow \mathsf{false}$

Figure 3. Reduction rules for conversion functions

$$\begin{aligned} \mathsf{tagof}(\sigma + \ldots + \sigma) &= \mathsf{union} \\ \mathsf{tagof}(\{\nu_1 : \sigma_1,\ \ldots\ , \nu_n : \sigma_n\}) &= \text{rec-anon} \\ \mathsf{tagof}(\nu\ \{\nu_1 : \sigma_1,\ \ldots\ , \nu_n : \sigma_n\}) &= \text{rec-named } \nu \\ \mathsf{tagof}(\hat{\sigma}\ \mathsf{option}) &= \mathsf{tagof}(\hat{\sigma}) \end{aligned}$$

The function is undefined for the $\top$ and null types, but this is not a problem because these types never need to be used as arguments in Figure 2. The $\top$ type is always eliminated using the (*top*) rule and the null type is eliminated using either (*null-1*) or (*null-2*).

The handling of unions is specified using three rules. When adding a non-union type to a union (*union-1*), the union may or may not contain a case with the same tag as the new type. If it does, the new type is combined with the existing one, otherwise a new case is added. When combining two unions (*union-2*), we group the cases that have a shared tags. Finally, the last rule (*union-3*) covers the case when we are combining two non-union types. As discussed earlier, union implicitly permits null values and so we use the following auxiliary function which makes nullable types non-nullable (when possible) to simplify the type:

$$\begin{aligned} \lfloor \hat{\sigma}\ \mathsf{option} \rfloor &= \hat{\sigma} \quad \text{(option)} \\ \lfloor \sigma \rfloor &= \sigma \quad \text{(otherwise)} \end{aligned}$$

The remaining rules are straightforward. For collections, we find the common supertype of its elements (*list*); for compatible primitive types, we choose one of the two (*prim*) and a common supertype with null is either the type itself (*null-1*) or an option (*null-2*).

Properties. The common supertype relation finds a single supertype (among multiple candidates). The following theorems specify that the found type is uniquely defined (up to reordering of fields in records and union cases) and is, indeed, a common supertype.

Theorem 1 (Uniqueness)**.** *If $\sigma_1 \triangledown \sigma_2 \vdash \sigma$ and $\sigma_1 \triangledown \sigma_2 \vdash \sigma'$ then it holds that $\sigma :> \sigma'$ and $\sigma' :> \sigma$.*

Proof. The assumptions and shapes of the types to be unified in the rules in Figure 2 are disjoint, with the exception of (*order*), (*sym*) and (*refl*). These rules always produce types that are subtypes of each other. This property is preserved by rules that use the common supertype relation recursively. □

Theorem 2 (Common supertype)**.** *If $\sigma_1 \triangledown \sigma_2 \vdash \sigma$ then it holds that $\sigma :> \sigma_1$ and $\sigma :> \sigma_2$.*

Proof. By induction over the common supertype derivation $\vdash$. □

We stated earlier that the common supertype relation minimises the use of union types (by preferring common numeric types or common record types when possible). This property can be stated and proved formally:

Theorem 3. *Given σ_1, σ_2 if there is s σ such that $\sigma :> \sigma_1$ and $\sigma :> \sigma_2$ and σ is not a union type, then $\sigma_1 \triangledown \sigma_2 \vdash \sigma'$ and σ' is not a union type.*

Proof. The only rule that introduces an union type is (*union-3*), which is used only when no other inference rule can be applied. □

4. Formalising type providers

In this section, we build the necessary theoretical framework for proving the relativised type safety property of F# Data type providers in Section 5. We start by discussing the runtime representation of structural values and runtime conversions (Section 4.1). Then we embed these into a formal model of an F# subset that is necessary for discussing type providers (Section 4.2) and we describe how F# Data type providers turn an inferred structural type into actual F# code (Section 4.3).

4.1 Structural values and conversions

In the model presented here, we represent JSON, XML and CSV documents using the same *structural value*[6]. Structural values are first-order and can be one of the primitive values (Boolean true or false, string t, integer i, decimal d and floating-point number f), a missing value (null), a collection and a record (named or unnamed):

$$\begin{aligned} s = {} & i \mid d \mid f \mid t \mid \mathsf{true} \mid \mathsf{false} \mid \mathsf{null} \\ & \mid [s_1; \ldots; s_n] \\ & \mid \{\nu_1 \mapsto s_1, \ldots, \nu_n \mapsto s_n\} \\ & \mid \nu\ \{\nu_1 \mapsto s_1, \ldots, \nu_n \mapsto s_n\} \end{aligned}$$

The first few cases represent primitive values (i for integers, d for decimals, f for floating point numbers and t for strings). A collection is written as a list of values in square brackets, separated by semicolons. A record can optionally start with its name ν, followed by a sequence of field assignments $\nu_i \mapsto s_i$ written in curly brackets.

[6] Here, we diverge from the actual implementation in F# Data which uses a different implementation for each format (reusing existing libraries).

As suggested by subtyping, our system permits certain runtime conversions (*e.g.* a numerical value 1 can be treated as true). These are captured by the following operations:

$$\begin{aligned} op = {} & \mathsf{asInt}(s) \mid \mathsf{asDec}(s) \mid \mathsf{asFloat}(s) \mid \mathsf{asBool}(s) \\ & \mid \mathsf{asStr}(s) \mid \mathsf{getChildren}(s) \mid \mathsf{getField}(\nu?, \nu, s) \\ & \mid \mathsf{isNull}(s) \mid \mathsf{isTag}(\mathsf{tag}, s) \end{aligned}$$

The reduction rules for these operations are defined in Figure 3. It is important to note that the relativised property we prove about F# Data specifies *sufficient*, but not *necessary* conditions. That is, a type provided based on a sample is guaranteed to work correctly on certain inputs, but may also handle several additional inputs.

The conversion functions follow this approach. They perform conversions required by the subtyping relation (*e.g.* integer is converted to decimal or a float) but also attempt additional conversions.

- **Widening operations.** The widening operations are required by the subtyping. These include converting to a wider numerical type (integer to decimal and float, decimal to float, converting 0 and 1 to Boolean).
- **Conversions.** Additional conversions are not strictly required, but prove useful in practice. These include converting to a smaller numerical type (such as float to integer) written as $i = \lfloor f \rfloor$, which is not defined in case of arithmetic overflow; conversion to and from string values written as "t represents i", which parses or formats a non-string value and also treating of decimals 0.0 and 1.0 as Booleans.
- **Structural operations.** Aside from primitives, the getChildren operation returns treats null as an empty collection. The getField function is defined separately for named and unnamed records (we write $\perp$ for a name of an unnamed record). Accessing a field of a named record checks that the expected record name matches the actual.
- **Helper functions.** The semantics also defines two helper functions. The isNull function is used to check whether value is null and isTag is used to test whether a value can be treated as a value of type associated with the specified tag (as defined in Section 3.3). The last line defines a "catch all" pattern, which returns false for all remaining cases.

4.2 Featherweight F#

The semantics fragment introduced in the previous section discusses values and operations that are used by F# Data. This section adds a minimal subset of the standard F# language.

We focus on the object system (which is what F# type providers produce) and so the following combines standard ML [14] features with Featherweight Java [10] classes. However, we only need classes with properties and without inheritance. A class has a single implicit constructor and the declaration closes over constructor parameters. To avoid including all of ML, we only select constructs for working with options and lists that we need later.

$$\begin{aligned} \tau = {} & \mathsf{int} \mid \mathsf{decimal} \mid \mathsf{float} \mid \mathsf{bool} \mid \mathsf{string} \\ & \mid C \mid \mathsf{StructVal} \mid \mathsf{list}\langle\tau\rangle \mid \mathsf{option}\langle\tau\rangle \end{aligned}$$

$$\begin{aligned} L &= \mathsf{type}\ C(\overline{x:\tau}) = \overline{M} \\ M &= \mathsf{member}\ N : \tau = e \end{aligned}$$

$$\begin{aligned} v = {} & s \mid \mathsf{None} \mid \mathsf{Some}(v) \mid \mathsf{new}\ C(\overline{v}) \mid [v_1; \ldots; v_n] \\ e = {} & op \mid e.N \mid \mathsf{new}\ C(\overline{e}) \mid \mathsf{if}\ e_1\ \mathsf{then}\ e_2\ \mathsf{else}\ e_3 \\ & \mid \mathsf{Some}(e) \mid \mathsf{match}\ e\ \mathsf{with}\ \mathsf{Some}(x) \to e_1 \mid \mathsf{None} \to e_2 \\ & \mid \mathsf{match}\ e\ \mathsf{with}\ [x_1; \ldots; x_n] \to e_1 \mid \_ \to e_2 \\ & \mid [e_1; \ldots; e_n] \mid \mathsf{List.map}\ (\lambda x \to e_1)\ e_2 \end{aligned}$$

In the above definition, τ denotes types including a type of all structural values s called StructVal. A class definition L consists of a constructor and zero or more members. Values v include previously defined structural values s and values for the option and list type; finally expressions e include previously defined operations op, class construction, member access, conditionals and expressions for working with option values and lists. We include List.map as a special construct to avoid making the language too complex.

$$\text{(member)}\quad \frac{\mathsf{type}\ C(\overline{x:\tau}) = \ldots\ \mathsf{member}\ N_i : \tau_i = e_i \ldots}{(\mathsf{new}\ C(\overline{v})).N_i \rightsquigarrow e_i[\overline{x} \leftarrow \overline{v}]}$$

$$\text{(ctx)}\quad E[e] \rightsquigarrow E[e'] \qquad (\text{when } e \rightsquigarrow e')$$

$$\text{(cond1)}\quad \mathsf{if\ true\ then}\ e_1\ \mathsf{else}\ e_2 \rightsquigarrow e_1$$

$$\text{(cond2)}\quad \mathsf{if\ false\ then}\ e_1\ \mathsf{else}\ e_2 \rightsquigarrow e_2$$

$$\text{(match1)}\quad \begin{array}{l}\mathsf{match\ None\ with}\\ \mathsf{Some}(x) \to e_1 \mid \mathsf{None} \to e_2\end{array} \rightsquigarrow e_2$$

$$\text{(match2)}\quad \begin{array}{l}\mathsf{match\ Some}(v)\ \mathsf{with}\\ \mathsf{Some}(x) \to e_1 \mid \mathsf{None} \to e_2\end{array} \rightsquigarrow e_1[x \leftarrow v]$$

$$\text{(match3)}\quad \begin{array}{l}\mathsf{match}\ [v_1; \ldots; v_n]\ \mathsf{with}\\ [x_1; \ldots; x_m] \to e_1 \mid \_ \to e_2\end{array} \rightsquigarrow e_2 \quad (m \neq n)$$

$$\text{(match4)}\quad \begin{array}{l}\mathsf{match}\ [v_1; \ldots; v_n]\ \mathsf{with}\\ [x_1; \ldots; x_n] \to e_1 \mid \_ \to e_2\end{array} \rightsquigarrow e_1[\overline{x} \leftarrow \overline{v}]$$

$$\text{(map)}\quad \mathsf{List.map}\ (\lambda x.e)\ [v_1; \ldots] \rightsquigarrow [e[x \leftarrow v_1]; \ldots]$$

Figure 4. Featherweight F# – Remaining reduction rules

Reduction. The reduction relation is of the form $e \rightsquigarrow e'$. We also write $e \rightsquigarrow^* e'$ to denote the reflexive and transitive closure of $\rightsquigarrow$. The reduction rules for primitive functions op were discussed earlier in Figure 3. The rules for other expressions defined above are provided in Figure 4.

The (*ctx*) rule performs a reduction inside a sub-expression specified by an evaluation context. This models the eager evaluation order of F#. An evaluation context E is defined as:

$$\begin{aligned} E = {} & f(E) \mid E.N \mid \mathsf{new}\ C(\overline{v}, E, \overline{e}) \mid \mathsf{if}\ E\ \mathsf{then}\ e_1\ \mathsf{else}\ e_2 \\ & \mid \mathsf{Some}(E) \mid \mathsf{match}\ E\ \mathsf{with}\ \mathsf{Some}(x) \to e_1 \mid \mathsf{None} \to e_2 \\ & \mid [\overline{v}; E; \overline{e}] \mid \mathsf{match}\ E\ \mathsf{with}\ [x_1; \ldots; x_n] \to e_1 \mid \_ \to e_2 \\ & \mid \mathsf{map}\ (\lambda x \to e)\ E \mid [\cdot] \end{aligned}$$

$$f \in \{\ \mathsf{asInt}, \mathsf{asDec}, \mathsf{asFloat}, \mathsf{asBool}, \mathsf{asStr}, \mathsf{getField}, \mathsf{getChildren}, \mathsf{isNull}, \mathsf{isTag}\ \}$$

The evaluation first reduces arguments of functions and the evaluation proceeds from left to right as denoted by $\overline{v}, E, \overline{e}$ in constructor arguments or $\overline{v}; E; \overline{e}$ in list initialization.

We write $e[\overline{x} \leftarrow \overline{v}]$ for the result of replacing variables $\overline{x}$ by values $\overline{v}$ in expression. The (*member*) rule reduces a member access using a class definition in the assumption to obtain the body of the expression. Finally, the remaining six rules provide standard reductions for conditionals and pattern matching.

The language is sufficient for our purpose, but simple. It does not provide recursion and so all expressions reduce to a value in a finite number of steps or get stuck due to an error condition. An error condition can be a wrong argument passed to conditional, pattern matching or one of the conversion functions from Figure 3.

$$\frac{}{L;\Gamma \vdash s : \mathsf{StructVal}} \qquad \frac{}{L;\Gamma \vdash n : \mathsf{int}}$$

$$\frac{L;\Gamma \vdash e : C \qquad \mathsf{type}\ C(\overline{x:\tau}) = \ldots \mathsf{member}\ N_i : \tau_i = e_i \ldots \in L}{L;\Gamma \vdash e.N_i : \tau_i}$$

$$\frac{L;\Gamma \vdash e_i : \tau_i \qquad \mathsf{type}\ C(x_1 : \tau_1, \ldots, x_n : \tau_n) = \ldots \in L}{L;\Gamma \vdash \mathsf{new}\ C(e_1, \ldots, e_n) : C}$$

Figure 5. Featherweight F# – Fragment of type checking

Type checking. The type checking rules in Figure 5 are written using a judgement $L;\Gamma \vdash e : \tau$ where the context also contains a set of class declarations L. The rules demonstrate the key difference between our language and standard ML or Featherweight Java:

- All structural values s have a type StructVal. Some of those have other types (primitive integers, decimals, floats, strings and Booleans as illustrated by the rule for n). For other values, StructVal is the only type – this includes records and null.
- A list containing other structural values $[s_1; \ldots; s_n]$ has a type StructVal. but can also have the list$\langle\tau\rangle$ type. Conversely, lists that contain non-structural values like objects or options are not of type StructVal.
- Primitive operations *op* are treated as functions, accepting StructVal and producing an appropriate type as the result.
- The two rules for checking class construction and member access are similar to corresponding rules of Featherweight Java.

An important part of Featherweight Java that is omitted here is the checking of type declarations (ensuring the bodies of members are well-typed). We omit this, because it is not needed for the relativized safety theorem, which considers only classes generated by our type provider (which are correct by construction).

4.3 Type providers

So far, we defined the type inference algorithm for structured data formats which produces a structural type σ based on one or more sample documents (Section 3). In Section 4.2, we defined a simplified model of F# language and evaluation, which also included conversions that model the F# Data library runtime (Section 4.3). In this section, we define how the type providers work, which links the two aspects discussed so far.

The type providers for XML, CSV and JSON all work in the same way. They take sample documents and infer a common supertype σ from the samples. From σ, they then generate F# types that are then exposed to the programmer[7].

Type provider mapping. When generating types, the type provider produces an F# type τ, expression that wraps a structural document value (of type StructVal) as a value of type τ and a collection of class definitions. We express it using the following mapping:

$$[\![-]\!]_e : \sigma \to \tau \times e' \times L$$

The semantics is parameterized by an expression e, which represents code to obtain a structural value that is being wrapped. Then, given an inferred structural type σ, the type provider produces an F# type τ, an expression e' which constructs a value of type τ using e and also a set of class definitions L.

Figure 6 shows the rules that define $[\![-]\!]_e$. Primitive types are all handled by a single rule. For a given structural type, it returns the corresponding F# type – the types are mapped directly with the exception of bit, which becomes an F# bool. The generated code calls an appropriate conversion function from Figure 3 on the input.

Handling of records is more interesting. We generate a new class C that takes a structural value as constructor parameter. For each record field, we generate a new member with the same name as the field[8] The body of the member calls getField and then passes this expression to $[\![\sigma_i]\!]$ which adds additional wrapping that maps the field (structural value of type σ_i) into an F# type τ_i. The returned expression creates a new instance of C and the mapping returns the class C together with all recursively generated definitions.

A collection type becomes an F# list$\langle\tau\rangle$. The returned expression calls getChildren (which turns null values into empty lists) and then uses List.map to convert all nested values to an F# type τ. The handling of option type is similar – it checks if the original value is null and if no, it wraps the recursively generated conversion expression e' in the Some constructor.

As discussed earlier, union types are also generated as classes with properties. Given a union type $\sigma_1 + \ldots + \sigma_n$, we get corresponding F# types τ_i and generate n members of type option$\langle\tau_i\rangle$. Each member return Some when the value is not null and has the right structure (checked by isTag). The type inference algorithm also guarantees that there is only one case for each type tag (Section 3.3) and there are no nested union types. Thus, checking for a tag is sufficient and we can also use the tag to identify the name of the generated member (using the nameof function).

Example 1. To illustrate how the type provision mechanism works, we now consider two simple examples. First, assume that the inferred type is a record with two fields (one optional) such as Person { Age : int option, Name : string }. Applying the rules from Figure 6 produces the following class:

```
type Person(v : StructVal) =
  member Age : option⟨int⟩ =
    if isNull (getField(Person, Age, v)) then None
    else Some(asInt(getField(Person, Age, v)))
  member Name : string =
    asStr(getField(Person, Name, v))
```

The body of the Age member is produced by the case for optional types applied to an expression getField(Person, Age, v). If the returned field is not null, then the member calls asInt (produced by the case for primitive types) and wraps the result in the Some constructor. Note that getField is defined even when the field does not exist, but returns null. This lets us treat missing fields as optional fields. The Name member is similar, but does not perform any checks. For completeness, a type corresponding to the record is Person and given a structural value s, we create a Person value by calling new Person(s).

Example 2. The second example illustrates the remaining parts of the type provision, including collections and union types. Reusing the Person type from the previous example, consider a collection [Person+string], which contains a mix of Person and string values.

```
type PersonOrString(v : StructVal) =
  member Person : option⟨Person⟩ =
    if isNull(v) then None else
    if isTag(rec-named Person, v) then
      Some(new Person(v)) else None
  member String : option⟨string⟩ =
    if isNull(v) then None else
    if isTag(string, v) then
      Some(asStr(v)) else None
```

[7] The actual implementation provides *erased types* as described in [23]. Here, we treat the code as actually generated. This is an acceptable simplification, because F# Data type providers do not rely on laziness that is available through erased types.

[8] The actual F# Data implementation also capitalizes the names.

$$[\![\sigma_p]\!]_e = \tau_p, op(e), \emptyset \quad \text{where}$$
$$\sigma_p, \tau_p, op \in \{\ (\mathsf{bit}, \mathsf{bool}, \mathsf{asBool}), (\mathsf{bool}, \mathsf{bool}, \mathsf{asBool})$$
$$(\mathsf{int}, \mathsf{int}, \mathsf{asInt}), (\mathsf{decimal}, \mathsf{decimal}, \mathsf{asDec}),$$
$$(\mathsf{float}, \mathsf{float}, \mathsf{asFloat}), (\mathsf{string}, \mathsf{string}, \mathsf{asStr})\ \}$$

$$[\![\nu?\ \{\nu_1 : \sigma_1, \ldots, \nu_n : \sigma_n\}]\!]_e =$$
$$C, \mathsf{new}\ C(e), L_1 \cup \ldots \cup L_n \cup \{L\} \quad \text{where}$$
$$C \text{ is a fresh class name}$$
$$L = \mathsf{type}\ C(v : \mathsf{StructVal}) \ = \ M_1 \ldots M_n$$
$$M_i = \mathsf{member}\ \nu_i : \tau_i = e_i$$
$$\tau_i, e_i, L_i = [\![\sigma_i]\!]_{e'}, \quad e' = \mathsf{getField}(\nu?, \nu_i, e)$$

$$[\![\ [\sigma]\ ]\!]_e = \mathsf{list}\langle\tau\rangle, \mathsf{List.map}\ (\lambda x \to e')\ (\mathsf{getChildren}(e)), L$$
$$\text{where}\ \ \tau, e', L = [\![\hat{\sigma}]\!]_x$$

$$[\![\hat{\sigma}\ \ \mathsf{option}]\!]_e =$$
$$\mathsf{option}\langle\tau\rangle, \mathsf{if}\ \mathsf{isNull}\ e\ \mathsf{then}\ \mathsf{None}\ \mathsf{else}\ \mathsf{Some}(e'), L$$
$$\text{where}\ \tau, e', L = [\![\hat{\sigma}]\!]_e$$

$$[\![\sigma_1 + \ldots + \sigma_n]\!]_e =$$
$$C, \mathsf{new}\ C(e), L_1 \cup \ldots \cup L_n \cup \{L\} \quad \text{where}$$
$$C \text{ is a fresh class name}$$
$$L = \mathsf{type}\ C(v : \mathsf{StructVal}) \ = \ M_1 \ldots M_n$$
$$M_i = \mathsf{member}\ \nu_i : \mathsf{option}\langle\tau_i\rangle =$$
$$\mathsf{if}\ \mathsf{isNull}(v)\ \mathsf{then}\ \mathsf{None}\ \mathsf{else}$$
$$\mathsf{if}\ \mathsf{isTag}(t_i, v)\ \mathsf{then}\ \mathsf{Some}(e_i)\ \mathsf{else}\ \mathsf{None}$$
$$\tau_i, e_i, L_i = [\![\sigma_i]\!]_e, \quad t_i = \mathsf{tagof}(\sigma_i), \quad \nu_i = \mathsf{nameof}(t_i)$$

$$[\![\top]\!]_v = [\![\mathsf{null}]\!]_v = \mathsf{StructVal}, v, \emptyset$$

$$\mathsf{nameof}(\mathsf{string}) = \mathsf{String} \qquad \mathsf{nameof}(\mathsf{number}) = \mathsf{Number} \qquad \mathsf{nameof}(\mathsf{rec\text{-}anon}) = \mathsf{Record}$$
$$\mathsf{nameof}(\mathsf{bool}) = \mathsf{Boolean} \qquad \mathsf{nameof}(\mathsf{collection}) = \mathsf{List} \qquad \mathsf{nameof}(\mathsf{rec\text{-}named}\ \nu) = \nu$$

Figure 6. Type provider – generation of featherweight F# types from inferred structural types

The type provider generates the above type and the collection type is mapped to an F# type $\mathsf{list}\langle\mathsf{PersonOrString}\rangle$. Given a structural document value s, the code to obtain the wrapped F# value is:

$$\mathsf{List.map}\ (\lambda x \to \mathsf{new}\ \mathsf{PersonOrString}(x))\ (\mathsf{getChildren}(s))$$

The PersonOrString type contains one property for each of the union case. In the body, they check that the value is not null and that it has the right structure (using the isTag function). This checks that the value is a record named Person or a string, respectively. If the conditions are satisfied, the value is converted to the F# type corresponding to the case and wrapped in Some.

4.4 Inferring types from values

Before concluding the section on formalizing type providers, there is one more missing piece. The common supertype algorithm in Section 3 discusses the core of the type inference for structural values, but we have not yet clarified how exactly it is used. We address this in the present section.

Given a JSON, XML or CSV document, the F# Data implementation constructs a structural value s from the sample (this is straightforward, but Section 6.1 discusses some interesting aspects). The following defines a mapping $\langle\!\langle - \rangle\!\rangle$ which turns a sample value s into a structural type σ:

$$\langle\!\langle 0 \rangle\!\rangle = \mathsf{bit} \qquad \langle\!\langle 1 \rangle\!\rangle = \mathsf{bit}$$
$$\langle\!\langle i \rangle\!\rangle = \mathsf{int} \qquad \langle\!\langle d \rangle\!\rangle = \mathsf{decimal}$$
$$\langle\!\langle f \rangle\!\rangle = \mathsf{float} \qquad \langle\!\langle t \rangle\!\rangle = \mathsf{string}$$
$$\langle\!\langle b \rangle\!\rangle = \mathsf{bool} \qquad \langle\!\langle \mathsf{null} \rangle\!\rangle = \mathsf{null}$$

$$\langle\!\langle [s_1; \ldots; s_n] \rangle\!\rangle = [\langle\!\langle s_1, \ldots, s_n \rangle\!\rangle]$$

$$\langle\!\langle \nu?\ \{\nu_1 \mapsto s_1, \ldots, \nu_n \mapsto s_n\} \rangle\!\rangle =$$
$$\nu?\ \{\nu_1 : \langle\!\langle s_1 \rangle\!\rangle, \ldots, \nu_n : \langle\!\langle s_n \rangle\!\rangle\}$$

$$\langle\!\langle s_1, \ldots, s_n \rangle\!\rangle = \sigma_n \quad \text{where}$$
$$\sigma_0 = \top,\ \forall i \in \{1..n\}.\ \sigma_{i-1} \triangledown \langle\!\langle s_i \rangle\!\rangle \vdash \sigma_i$$

Primitive values are mapped to their corresponding types, with the exception of 0 and 1, which are inferred as bit. For records, we return a type with field types inferred based on individual values.

The interesting part is inferring type based on multiple samples. We overload the notation and write $\langle\!\langle s_1, \ldots, s_n \rangle\!\rangle$ for a type inferred from multiple samples. This uses the common supertype relation to find a common type for all values (starting with $\top$). This operation is used at the top-level (when calling type provider with multiple samples) and also when inferring the type of a collection.

5. Relativized type safety

Informally, the safety property of F# Data type providers states that, given representative sample documents, any code that can be written using the provided types is guaranteed to work. We call this *relativized safety*, because we cannot avoid *all* errors. In particular, the user can always use the type provider with an input that has a different structure than any of the samples – and in this case, it is expected that the code will fail at runtime (throw an exception in the actual implementation or by getting stuck in our model).

The key question is, what is a representative sample? Given a set of sample documents, the provided type is guaranteed to work if the inferred type of the input is a subtype of any of the sample documents. Going back to Section 3.2, this means that:

- Input can contain smaller numerical values (for example, if a sample contains float, the input can contain an integer).
- Records in the actual input can have additional fields
- Records in the actual input can have fewer fields, provided that the type of the fields is marked as optional in the sample
- Union types in the input can have both fewer or more cases
- When we have a union type in the sample, the actual input can also contain just values of one of the union cases

The following lemma states that the provided code (generated in Figure 6) works correctly on inputs s' that is a subtype of the sample s. More formally, we require that the provided expression (using s' as the input) can be reduced to a value and, if it is a class, all its members can also be reduced to values.

Lemma 1 (Correctness of provided types)**.** *Given a sample value s and an input value s' such that $\langle\!\langle s \rangle\!\rangle :> \langle\!\langle s' \rangle\!\rangle$ and provided type, expression and class declarations $\tau, e, L = [\![\langle\!\langle s \rangle\!\rangle]\!]_{s'}$, then $e \rightsquigarrow^* v$ and if τ is a class ($\tau = C$) then for all members N_i of the class C, it holds that $e.N_i \rightsquigarrow^* v$.*

Proof. By induction over the structure of $[\![-]\!]_e$. For primitives, the conversion functions accept all subtypes. For other cases, analyse the the provided code to see that it can work on all subtypes (*e.g.* getChildren works on null values, getField returns null when a field is missing and, when isTag returns true, then a value has the structure required by the corresponding case). □

Now that we know that the provided types are correct with respect to the subtyping relation, we can look at the main theorem of the

paper. It states that, for any input (which is a subtype of any of the samples) and any expression e, a well-typed program that uses the provided types does not "go wrong".

Using the standard approach to syntactic type safety [25], we prove the type preservation (reduction does not change type) and progress (an expression that is not a value can be reduced).

Theorem 4 (Relativized safety)**.** *Assume $s_1, \ldots, s_n$ are samples, $\sigma = \langle\!\langle s_1, \ldots, s_n \rangle\!\rangle$ is an inferred type and $\tau, e, L = [\![\sigma]\!]_x$ are a type, expression and class definitions generated by a type provider.*

Then for all new inputs s' such that $\exists i.(\langle\!\langle s_i \rangle\!\rangle :> \langle\!\langle s' \rangle\!\rangle)$, let $e_s = e[x \leftarrow s']$ be an expression (of type τ) that wraps the input in a provided type. Then, for any expression e_c (user code) such that $\emptyset; y : \tau \vdash e_c : \tau'$ and that does not contain any structural values as sub-expressions, it is the case that $e_c[y \leftarrow e_s] \rightsquigarrow^* v$ for some value v and also $\emptyset; \vdash v : \tau$.

Proof. We discuss the two parts of the proof separately as type preservation (Lemma 2) and progress (Lemma 3). □

Lemma 2 (Preservation)**.** *Given the class definitions L generated by a type provider as specified in the assumptions of Theorem 4, then if $\Gamma \vdash e : \tau$ and $e \rightsquigarrow^* e'$ then $\Gamma \vdash e' : \tau$.*

Proof. By induction over the reduction $\rightsquigarrow$. The cases for the ML subset of Featherweight F# are standard. For (*member*), we check that code generated by type providers in Figure 6 is well-typed. □

The progress lemma states that evaluation of a well-typed program does not reach an undefined state. This is not a problem for the ML subset and object-oriented subset of the calculus. The problematic part are the conversion functions (Figure 3). Given a structural value (which has a type StructVal in our language), the reduction can get stuck if the value does not have a structure required by a specific conversion function used.

The Lemma 1 guarantees that this does not happen. In Theorem 4, we carefully state that we only consider expressions e_c which "[do] not contain any structural values as sub-expressions". This makes sure that the only code working with structural values is the code generated by the type provider.

Lemma 3 (Progress)**.** *Given the assumptions and definitions from Theorem 4, it is the case that $e_c[y \leftarrow e_s] \rightsquigarrow^* v$.*

Proof. By induction over the typing derivation of $L; \emptyset \vdash e_c[y \leftarrow e_s] : \tau'$. The cases for the ML subset are standard. For member access, we rely on Lemma 1. □

5.1 Discussion

The *relativized safety* property does not guarantee the same amount of safety as standard type safety for programming languages without type providers. However, it reflects the reality of programming with external data sources that is increasingly important in the age of web [15]. So, type providers do not reduce the safety – they simply make the existing issues visible.

The actual implementation in F# Data throws an exception for invalid inputs. In contrast, the calculus presented here simply gets stuck. We could extend the calculus with exceptions, but that would obscure the purpose – precisely specifying when the code using type providers does not "go wrong."

As mentioned earlier, *relativized safety* specifies a sufficient, but not a necessary condition. This is also reflected in the conversion functions which allow additional conversions not required by the subtyping relation. In practice, the additional flexibility proves useful (as minor variations in input formats are common). However, an interesting stricter alternative would be to check that an input has the required type when it is loaded – and throw an exception immediately when loading data rather than on member access.

Finally, the formal model presented here ignores two aspects of the real type provider mechanisms. As discussed here, we do not use the *type erasure* mechanism, but treat type providers as if they were actually generating code. For XML, CSV and JSON, both models can be used. However, one interesting aspect of erasing type providers is that the types can be generated lazily – only classes that are actually used in the type checked code are generated. Extending our formalism to capture this aspect is an interesting future work relevant to other type providers – including the World Bank type provider that is also available in the F# Data library.

6. Implementation

The theoretical model in Section 4 presents the core ideas behind the type providers for structured data formats. However, an important part of the success of the F# Data library is also its pragmatic approach, which makes it suitable for use with real-world data.

The formal model discusses some of the pragmatic choices, such as the preference for records over unions and the ability to treat 0 and 1 as both Booleans and numbers. In this section, we briefly discuss some of the remaining practical concerns, starting with the handling of collections.

6.1 Parsing structured data

In this paper, we treat the XML, JSON and CSV formats uniformly as *structural values*. The definition of structural values is, indeed, rich enough to capture all three formats. The structure is similar to JSON (it has primitive values, records and collections). We also added *named* records, which are needed for XML.

- When reading JSON, we directly create a corresponding structural value (with all records unnamed). Optionally, we also read numerical values (and dates) stored in strings.
- When reading CSV, we read each row as an unnamed record and return them in a collection. Optionally, the read values are parsed as numbers or Booleans and missing values become null.
- When reading XML, we create a named record for each node. Attributes become record fields and body becomes a special field named •

The reading of XML documents is perhaps the most interesting. To demonstrate how the body is treated, consider the following:

```
<root id="1">
  <item>Hello!</item>
</root>
```

This XML becomes a record named root with fields id and •. The nested element contains only the • field containing the inner text:

$$\mathsf{root}\ \{\mathsf{id} \mapsto 1, \bullet \mapsto [\mathsf{item}\ \{\bullet \mapsto \texttt{"Hello!"}\}]\}$$

When generating types for types inferred from XML, we also include a special case to remove the • node using the fact that this can be only a collection (of elements) or a primitive value.

Finally, the XML type provider in F# Data includes an additional option to use *global inference*. In that case, the inference algorithm from Section 4.4 is replaced with an alternative that unifies the types of all records with the same name. This is useful when the sample is, for example, an XHTML document – in this case, all occurrences of an element (*e.g.* <a>) are treated as the same type.

6.2 Heterogeneous collections

When introducing the XML type provider (Section 2.3), we briefly mentioned that F# Data implements a special handling of heterogeneous collections. In the example, the provider generated a type

$$\text{(col)}\ \frac{\mathsf{tagof}(\sigma_i) = \mathsf{tagof}(\sigma'_i) \quad \sigma_i \triangledown \sigma'_i \vdash \sigma''_i \quad \phi_i \triangledown \phi'_i \vdash \phi''_i \qquad (\forall i \in \{1..k\})}{[\sigma_1, \psi_1| \ldots |\sigma_k, \psi_k| \ldots |\sigma_n, \psi_n] \triangledown [\sigma'_1, \psi'_1| \ldots |\sigma'_k, \psi'_k| \ldots |\sigma'_m, \psi'_m] \vdash [\sigma''_1, \psi''_1| \ldots |\sigma''_k, \psi''_k|\sigma_{k+1}, \lceil\psi_{k+1}\rceil| \ldots |\sigma_n, \lceil\psi_n\rceil|\sigma'_{k+1}, \lceil\psi'_{k+1}\rceil| \ldots |\sigma'_m, \lceil\psi'_m\rceil]}$$

$$\lceil\phi\rceil = \phi' \text{ such that } \phi \triangledown 1? \vdash \phi'$$
$$\phi \triangledown \phi' \vdash \phi \quad \text{when} \quad \phi :> \phi'$$
$$\phi \triangledown \phi' \vdash \phi' \quad \text{when} \quad \phi' :> \phi$$

Figure 7. Inference judgement that defines the common supertype of two heterogeneous collections

with Title member of type string (corresponding to a nested element `<title>` that appears exactly once) and Items member (returning a list of values obtained from multiple `<item>` elements).

To capture this behaviour, we need to extend our earlier treatment of collections. Rather than storing a single type for the elements as in $[\sigma]$, we store multiple possible element types. However, this is not just a union type as we also store *inferred multiplicity* of elements for each of the types:

$$\begin{array}{l} \psi = 1 \mid 1? \mid * \\ \sigma = \ \ldots \mid [\sigma_1, \psi_1| \ldots |\sigma_n, \psi_n] \end{array}$$

The multiplicities 1, $1?$ and $*$ represent *exactly one*, *zero or one* and *zero or more*, respectively. Thus, the inferred type of the collection in the RSS feed would be $[\mathsf{title}\ \{\ldots\}\ , 1|\ \mathsf{item}\ \{\ldots\}, *]$, which reads as a collection containing exactly one title record and any number of item records.

The subtyping relation on heterogeneous collections specifies that a subtype can have fewer cases provided that they do not have to appear exactly once. A subtype can also have a stricter specification of multiplicity (the ordering is $* :> 1? :> 1$).

Finding a common supertype of heterogeneous collections is analogous to the handling of union types. The key rule is (*col*) in Figure 7. It merges cases with the same tag (by merging both the type and the multiplicity). For cases that appear only in one collection, we ensure that the multiplicity is *zero or one* or *zero or more* using the auxiliary definition $\lceil - \rceil$.

6.3 Inferring and generating unions

The F# Data type providers prefer types that do not contain unions (Theorem 3). The motivation for this is twofold. First, the current support for type providers in the F# compiler does not allow type providers to generate F#-specific types (including discriminated unions). This means that the provided type cannot use idiomatic F# representation. Second, when accessing data, using records aids the explorability – when a value is a record, users can type "." and most modern F# editors provide auto-completion list with members.

This choice has an important desirable consequence – it allows treating a union with additional cases as a subtype of another union and, hence, the provided code is guaranteed to work on *more* inputs. Consider a type that is either a Person record or just a string. The type provider exposes it as the following F# class:

```
type PersonOrString =
  member Person : option⟨Person⟩
  member String : option⟨string⟩
```

If we provided an algebraic data type (discriminated union), the consumer would have to use pattern matching that would cover *two* cases. The above type forces the user to also handle a third case, when both properties return None. This also means that union types can silently skip over null values.

That said, exposing a discriminated union (perhaps together with better tooling) is certainly an attractive alternative that we plan to consider in the future. This can also be combined with type inference that chooses between records and union types based on some statistical metric (see related work in Section 7).

6.4 Practical experiences

An important concern about using F# Data in practice is the handling of schema change. When using a type provider, the sample is captured at compile-time. If the schema changes later (so that the actual input is no longer a subtype of the samples), the program using the type provider fails at run-time and it is the developer's responsibility to handle the exception. However, this is the same problem that happens when reading data using any other library.

F# Data can help discover such errors earlier. The example in Section 1 points the JSON type provider to a sample using a live URL. This has the advantage that a re-compilation fails when the schema changes, which is an indication that the program needs to be updated to reflect the change. If this is undesirable, it is always possible to cache the sample locally.

In general, there is no better solution for plain XML, CSV and JSON data sources. Some data sources provide versioning support (with meta-data about how the schema changed). For those, a type provider could adapt automatically, but we leave this for future.

An approach that we find useful in practice is to keep a set of samples. When the program fails at run-time (because a service returned data in an unexpected format), the new input can be added as an additional sample, which then shows what parts of code need to be modified.

7. Related and future work

The F# Data library connects two lines of research that have been previously disconnected. The first is extending the type systems of programming languages to accommodate external data sources and the second is inferring types for real-world data sources.

The type provider mechanism has been introduced in F# [22, 23], added to Agda [3] and used in areas such as semantic web [17]. Our paper is novel in that it shows the programming language theory behind a concrete type providers.

Extending the type systems. A number of systems integrate external data formats into a programming language. Those include XML [9, 20] and databases [5]. In both of these, the system either requires the user to explicitly define the schema (using the host language) or it has an ad-hoc extension that reads the schema (*e.g.* from a database). LINQ [13] is more general, but relies on code generation when importing the schema.

The work that is most similar to F# Data is the XML and SQL integration in Cω [12]. It extends a C# like object-oriented language with types similar to our structural types (including nullable types, choices with subtyping and heterogeneous collections with multiplicities). However, Cω does not infer the types from samples and extends the type system of the host language (rather than using a general purpose embedding mechanism).

Advanced type systems. Aside from type providers, a number of other advanced type system features could be used to tackle the problem discussed in this paper. The Ur [2] language has a rich system for working with records; meta-programming [18], [6] and multi-stage programming [24] could be used to generate code for the provided types. However, as far as we are aware, none of these systems have been used to provide the same level of integration with XML, CSV and JSON.

Another approach would be to use gradual typing [19, 21] and add types for structured data formats to existing dynamic language to check, for example, JSON manipulation in JavaScript.

Typing real-world data. The second line of research related to our work focuses on inferring structure of real-world data sets. A recent work on JSON [4] infers a succinct type using MapReduce to handle large number of samples. It fuses similar types based on a type similarity measure. This is more sophisticated than our technique, but it would make formally specifying the safety properties (Theorem 4) difficult. Extending our *relativized safety* property to a *probabilistic safety* is an interesting future work.

The PADS project [7, 11] tackles a more general problem of handling *any* data format. The schema definitions in PADS are similar to our structural type. The structure inference for LearnPADS [8] infers the data format from a flat input stream. A PADS type provider could follow many of the patterns we explore in this paper, but formally specifying the safety property would be challenging.

8. Conclusions

In this paper, we explored the F# Data type providers for structured data formats such as XML, CSV and JSON. As most real-world data do not come with an explicit schema, the library uses *type inference* that deduces a type from a set of samples. Next, it uses the *type provider* mechanism to integrate the inferred type directly into the F# type system.

The type inference algorithm we use is based on a common supertype relation. For usability reasons, the algorithm attempts to infer type that does not contain unions and prefers records with optional fields instead. The algorithm is simple and predictable, which is important as developers need to understand how changing the samples affects the resulting types.

In the second part of the paper, we explore the programming language theory behind type providers. F# Data is a prime example of type providers, but our work also demonstrates a more general point. The types generated by type providers can depend on external input (such as samples) and so we can only formulate *relativized safety property*, which says that a program is safe only if the actual inputs satisfy additional conditions – in our case, they have to be subtypes of one of the samples.

The type provider mechanism has been described before, but this paper is novel in that it explores concrete type providers from the perspective of programming language theory. This is particularly important as the F# Data library is becoming de-facto standard tool for data access in F#[9] and other languages are beginning to adopt mechanisms similar to F# type providers.

Acknowledgments

We would like to thank to the F# Data contributors on GitHub and other colleagues working on type providers, including Jomo Fisher, Keith Battocchi and Kenji Takeda.

References

[1] G. Bracha, M. Odersky, D. Stoutamire, and P. Wadler. Making the future safe for the past: Adding genericity to the java programming language. *Acm sigplan notices*, 33(10):183–200, 1998.

[2] A. Chlipala. Ur: Statically-typed metaprogramming with type-level record computation. In *ACM Sigplan Notices*, volume 45, pages 122–133. ACM, 2010.

[3] D. R. Christiansen. Dependent type providers. In *Proceedings of Workshop on Generic Programming*, WGP '13, pages 25–34, 2013. ISBN 978-1-4503-2389-5.

[4] D. Colazzo, G. Ghelli, and C. Sartiani. Typing massive json datasets. In *International Workshop on Cross-model Language Design and Implementation*, XLDI '12, 2012.

[5] E. Cooper, S. Lindley, P. Wadler, and J. Yallop. Links: Web programming without tiers. In *Formal Methods for Components and Objects*, pages 266–296. Springer, 2007.

[6] J. Donham and N. Pouillard. Camlp4 and Template Haskell. In *Commercial Users of Functional Programming*, 2010.

[7] K. Fisher and R. Gruber. PADS: a domain-specific language for processing ad hoc data. *ACM Sigplan Notices*, 40(6):295–304, 2005.

[8] K. Fisher, D. Walker, and K. Q. Zhu. LearnPADS: Automatic tool generation from ad hoc data. In *Proceedings of International Conference on Management of Data*, SIGMOD '08, pages 1299–1302, 2008.

[9] H. Hosoya and B. C. Pierce. XDuce: A statically typed xml processing language. *Transactions on Internet Technology*, 3(2):117–148, 2003.

[10] A. Igarashi, B. Pierce, and P. Wadler. Featherweight java: A minimal core calculus for java and gj. In *ACM SIGPLAN Notices*, volume 34, pages 132–146. ACM, 1999.

[11] Y. Mandelbaum, K. Fisher, D. Walker, M. Fernandez, and A. Gleyzer. PADS/ML: A functional data description language. In *ACM SIGPLAN Notices*, volume 42, pages 77–83. ACM, 2007.

[12] E. Meijer, W. Schulte, and G. Bierman. Unifying tables, objects, and documents. In *Workshop on Declarative Programming in the Context of Object-Oriented Languages*, pages 145–166, 2003.

[13] E. Meijer, B. Beckman, and G. Bierman. LINQ: Reconciling object, relations and XML in the .NET Framework. In *Proceedings of the International Conference on Management of Data*, SIGMOD '06, pages 706–706, 2006.

[14] R. Milner. *The definition of standard ML: revised*. MIT press, 1997.

[15] T. Petricek and D. Syme. In the age of web: Typed functional-first programming revisited. *Submitted to post-proceedings of ML Workshop*, 2014.

[16] T. Petricek, G. Guerra, and Contributors. F# Data: Library for data access, 2015. URL http://fsharp.github.io/FSharp.Data/.

[17] S. Scheglmann, R. Lämmel, M. Leinberger, S. Staab, M. Thimm, and E. Viegas. Ide integrated rdf exploration, access and rdf-based code typing with liteq. In *The Semantic Web: ESWC 2014 Satellite Events*, pages 505–510. Springer, 2014.

[18] T. Sheard and S. P. Jones. Template meta-programming for haskell. In *Proceedings of the 2002 ACM SIGPLAN workshop on Haskell*, pages 1–16. ACM, 2002.

[19] J. G. Siek and W. Taha. Gradual typing for functional languages. In *Scheme and Functional Programming Workshop*, pages 81–92, 2006.

[20] M. Sulzmann and K. Z. M. Lu. A type-safe embedding of XDuce into ML. *Electr. Notes in Theoretical Comp. Sci.*, 148(2):239–264, 2006.

[21] N. Swamy, C. Fournet, A. Rastogi, K. Bhargavan, J. Chen, P.-Y. Strub, and G. Bierman. Gradual typing embedded securely in javascript. In *ACM SIGPLAN Notices*, volume 49, pages 425–437. ACM, 2014.

[22] D. Syme, K. Battocchi, K. Takeda, D. Malayeri, and T. Petricek. Themes in information-rich functional programming for internet-scale data sources. In *Proceedings of the Workshop on Data Driven Functional Programming*, DDFP'13.

[23] D. Syme, K. Battocchi, K. Takeda, D. Malayeri, J. Fisher, J. Hu, T. Liu, B. McNamara, D. Quirk, M. Taveggia, W. Chae, U. Matsveyeu, and T. Petricek. Strongly-typed language support for internet-scale information sources. Technical Report MSR-TR-2012-101, Microsoft Research, September 2012.

[24] W. Taha and T. Sheard. Multi-stage programming with explicit annotations. *SIGPLAN Not.*, 32(12):203–217, 1997. ISSN 0362-1340.

[25] A. K. Wright and M. Felleisen. A syntactic approach to type soundness. *Information and computation*, 115(1):38–94, 1994.

[9] At the time of writing, the library has 37,000 downloads on NuGet; over 1600 commits and 36 contributors on GitHub.