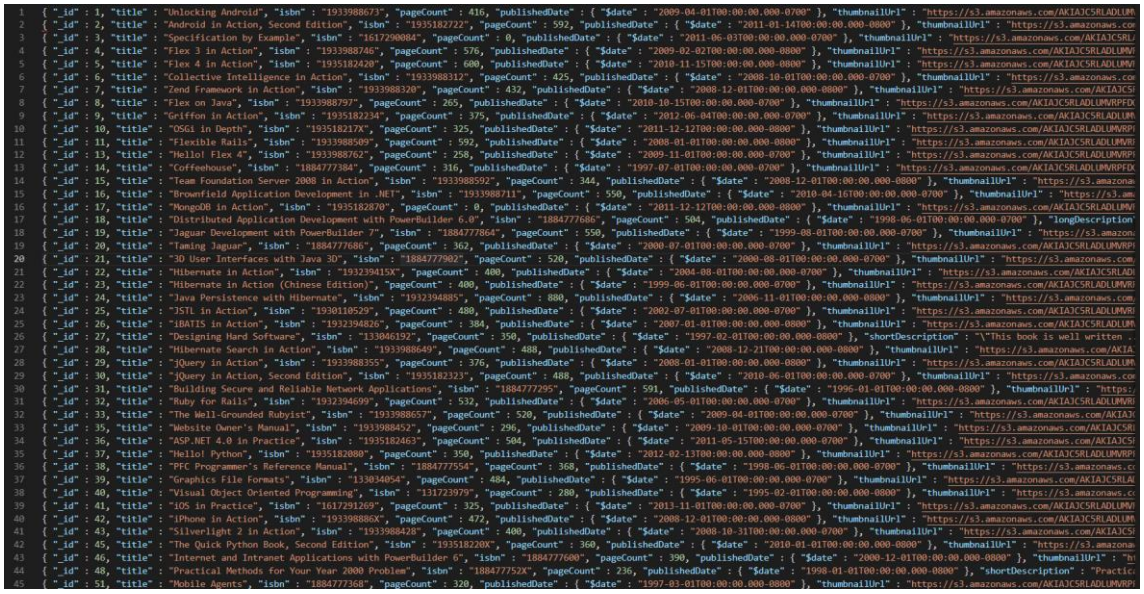
ANÁLISIS DE UN DATASET

Para la práctica de JSON y JSON Schema se pide inspeccionar un dataset de Kaggle o de Github y escribir un mini-reporte analizándolo.

De entre las opciones existentes, se ha escogido el dataset “books.json” del repositorio GitHub https://github.com/ozlerhakan/mongodb-json-files.git . El dataset en cuestión consta de 431 libros del mundo de la informática y la programación.

Una imagen en crudo de este fichero .json se puede ver a continuación:

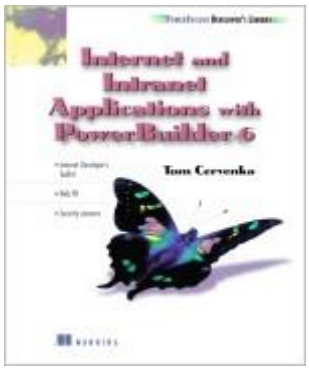


Para poder realizar el análisis es necesario tener estos datos en un formato más legible y, por ello, se ha escogido un libro en concreto y se han adaptado sus propiedades:

```
{
    "_id" : 46,
    "title" : "Internet and Intranet Applications with PowerBuilder 6",
    "isbn" : "1884777600",
    "pageCount" : 390,
    "publishedDate" : { "$date" : "2000-12-01T00:00:00.000-0800" },
    "thumbnailUrl" : "https://s3.amazonaws.com/AKIAJC5RLADLUMVRPFDQ.book-thumb-images/cervenka.jpg",
    "longDescription" : "If you're a PowerBuilder programmer, Internet and Intranet Applications with PowerBuilder 6 is your ticket to learning Web.
    PB and related technologies. The book covers everything you need to know to build web browser and server programs with the PowerBuilder 6 Internet Toolkit.
    Also covered is how to write winsock programs with PB, and Distributed PB is covered to the extent necessary to learn Web.PB.",
    "status" : "PUBLISH",
    "authors" : [ "Tom Cervenka" ],
    "categories" : [ "PowerBuilder" ]
}
```

Como se puede observar, las propiedades de cada libro cubren las características típicas, desde el título y los autores hasta el ISBN. Además, cuenta con un identificador unívoco para cada libro “_id”.

Los tipos de cada una de ellas son variados: hay strings, numbers, arrays y colecciones. Un ejemplo de colección sería la propiedad “publishDate”, la cuál es definida mediante una clave-valor como es la variable “$date” y la fecha en formato string. Por otro lado, la propiedad “thumbnailUrl” tiene como tipo un string que abstrae la URL de la imagen de la portada del libro.



Por último, cabe destacar los tipos arrays en propiedades como “authors” o “categories”, pues al poder tener varios valores es adecuado definirlas así. Concretamente en este ejemplo el libro solo tiene una categoría y un autor, pero en varios de los otros libros del dataset sí que tienen más.

Analizando el resto de los libros del conjunto se puede ver que en muchos de ellos está, como en el analizado en este reporte, la propiedad “longDescription”; en otros, en vez de esta, aparece la propiedad “shortDescription”; y, en otros, no aparece ninguna de las dos mencionadas. Esto lleva a concluir que en el esquema de este JSON esa propiedad en concreto no aparecerá como requerida mientras que el resto si lo estarán.