CSVファイル読み込みのデバッグに1日かかった件

2024/05/29
第164回PHP勉強会＠東京

https://fullcustomize.com
古家康裕（ふるややすひろ）

自己紹介

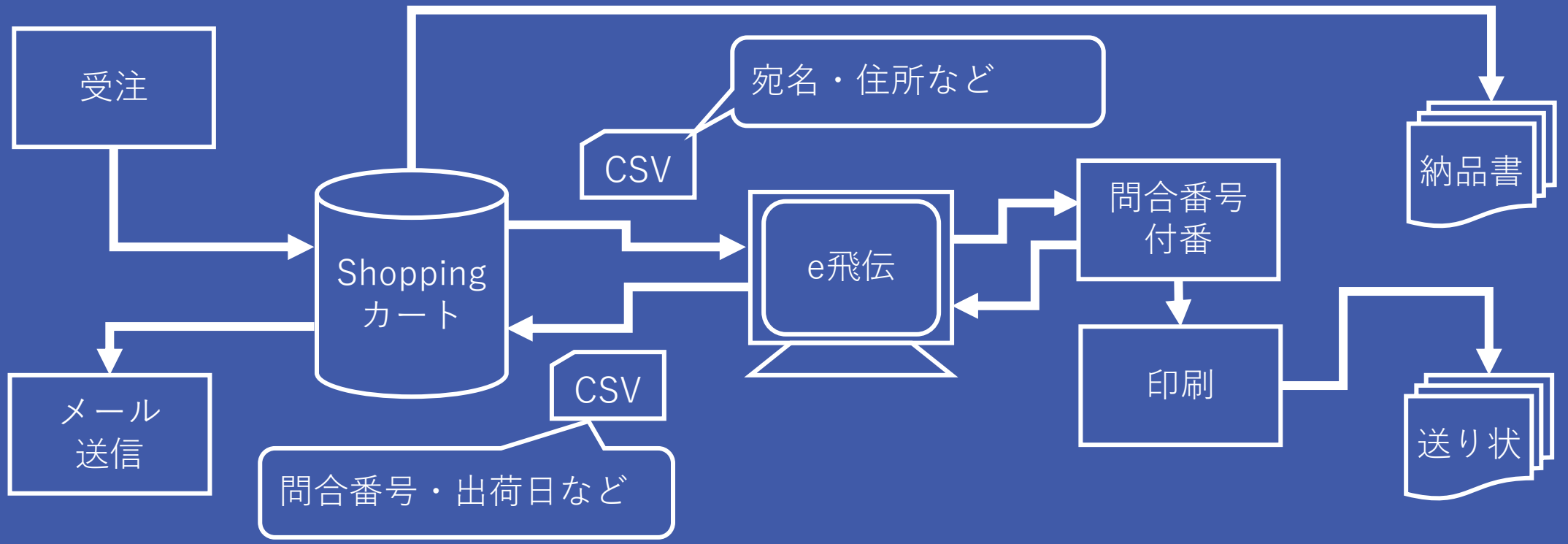
- 古家康裕（ふるややすひろ）PHP歴：約25年

- 『全て自社開発なのでフルカスタマイズOK』

- リストラ４回　→　会社員不向き→独立して自営
- 多重下請構造がイヤ　→　クライアントから直接受注
- XAMPP＋自社FW　→　維持工数の軽減
- サブスク契約　→　Win-Win

e飛伝IIサービス終了（2024年3月31日）。

佐川急便の送り状作成ソフト



e飛伝III

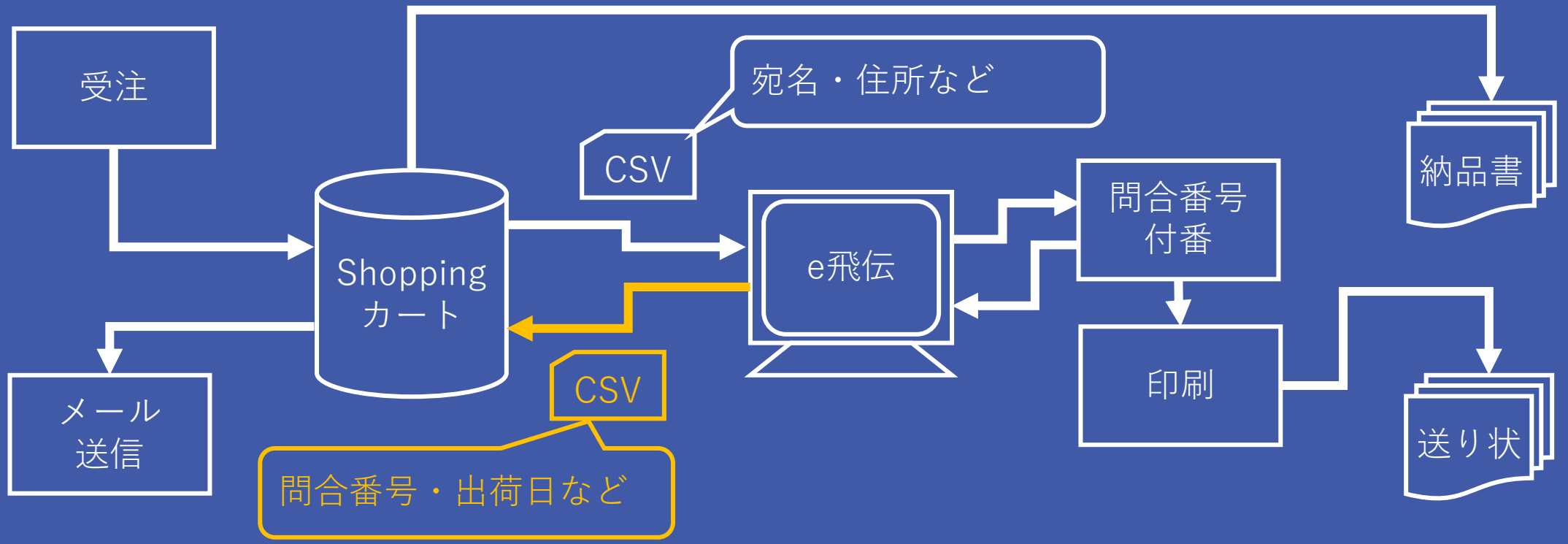
特徴：

- 新サービスへの対応
- オンライン版

技術的変更：

- CSVレイアウト変更
- 文字コードがUTF-8に

バグ：問合せ番号の「“”」が取れない



var_dump($csv_decoded);

```
array(){
    [0]=>string(8) ””123456”” // お問い合わせ番号
    [1]=>string(8) “123-4567” // 郵便番号
    [2]=>string(20) “東京都港区・・・・” // 住所
    [3]=>string(10) “2024/5/29” // 出荷日
            .
            .
            .
}
```

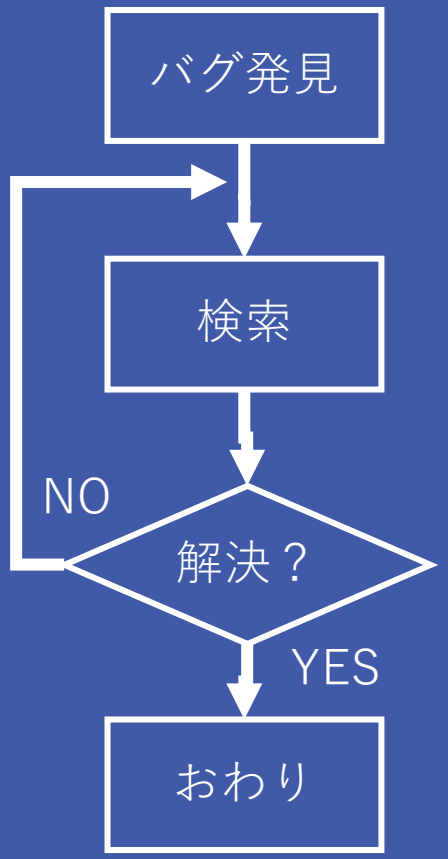
原因：CSVがBOMありUTF-8だった

BOM（Byte Order Mark）は、テキストデータの先頭に配置される特殊なバイト列のことです。主にUnicodeテキストファイルで使用され、ファイルがどのエンディアン（バイトオーダー）でエンコードされているかを示します。

UTF-8のBOMのバイトパターン：

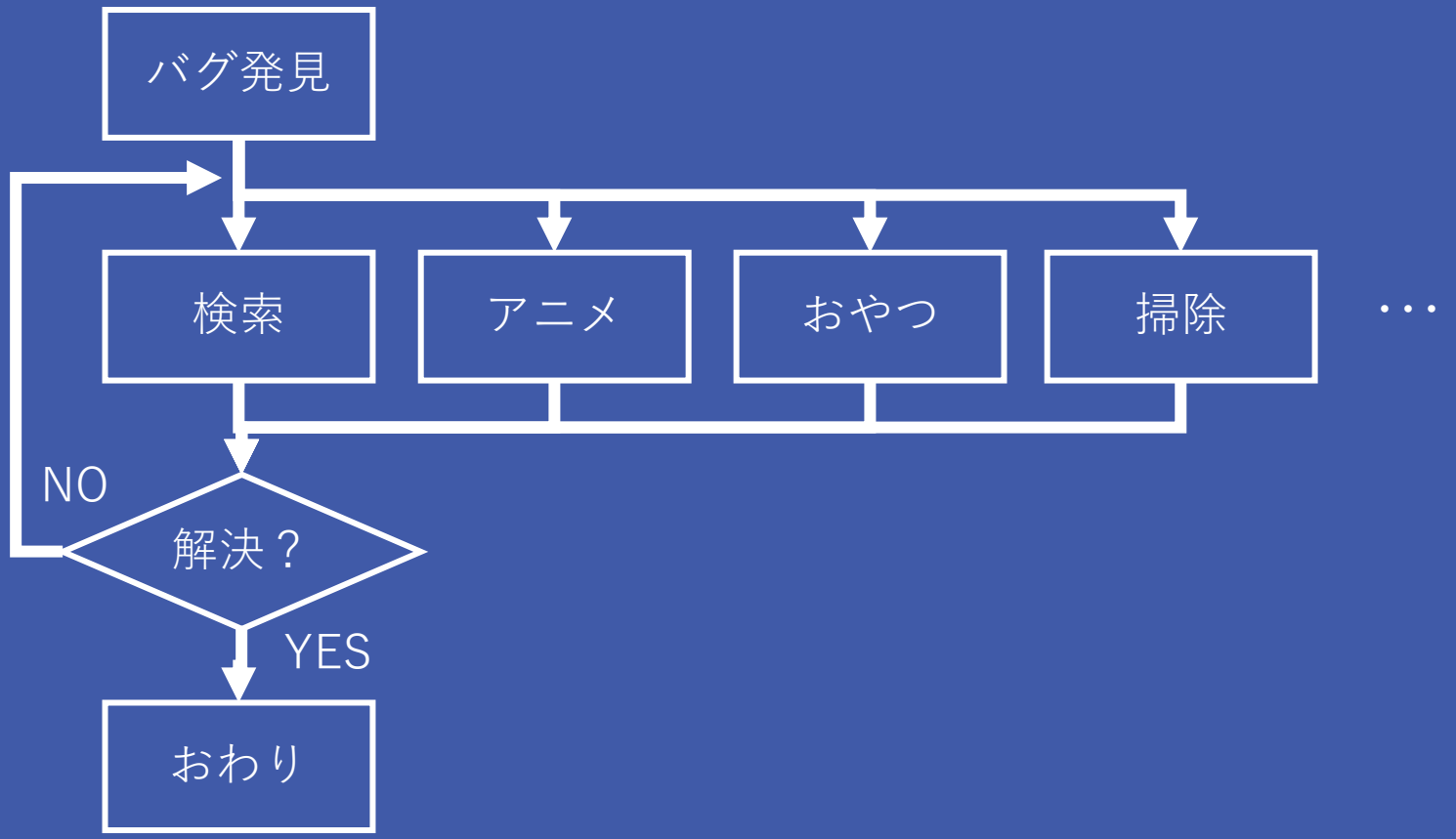
EF BB BF

デバッグに1日かかった件



…

デバッグに1日かかった件



どうすれば良かったのか（良いのか）？

Chat GPTに訊いてみた（１）

CSVをデコードするときに、クオーテーションマークが外れない問題の原因を挙げてください。

1. 不正なエスケープ処理：
2. 不適切なクオートの使用：
3. エンコーディングの不一致：デコード時に適切なエンコーディングが指定されていない場合、文字化けやデータの歪みが発生し、クオーテーションマークが外れる可能性があります。
4. CSVの仕様に基づく不整合：

Chat GPTに訊いてみた（２）

３．について詳細を質問します。データがUTF-8でエンコーディングされている場合の原因を挙げてください。

1. BOM（Byte Order Mark）の有無：UTF-8でエンコードされたCSVファイルの先頭にBOM（Byte Order Mark）が含まれていない場合、一部のCSVパーサーはエンコーディングを正しく認識せず、文字化けやクオーテーションマークが外れる問題が発生する可能性があります。
2. エンコーディングの不一致：
3. 不正な文字の使用：

Chat GPTに訊いてみた（３）

PHPでBOM付きUTF-8のCSVをデコードする
ソースコードのサンプルを提供してください。

```
<?php
if ($_SERVER['REQUEST_METHOD'] == 'POST' &&
isset($_FILES['csvFile'])) {
    $fileTmpPath = $_FILES['csvFile']['tmp_name'];
    $fileName = $_FILES['csvFile']['name'];
    $fileSize = $_FILES['csvFile']['size'];
    $fileType = $_FILES['csvFile']['type'];

（以下省略）
```

結論

Chat GPTれ！
（じぴれ！）