



COMP704 Research and Development Project

**VN01** 3D acupuncture healthcare data management and treatment system

# Data Integration Process

**Supervisor:**

Dr Nhan Le Thi

**Team Members:**

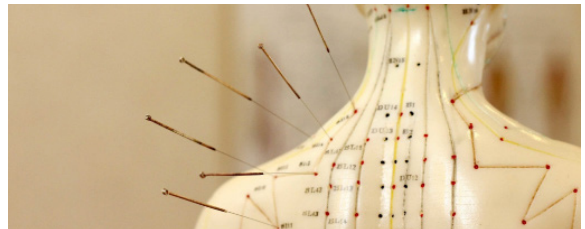
21142643	Chuong Pham Dinh
21142377	Nhan Nguyen Cao
21142355	Tan Le Tran Ba
21142358	Trang Ho Ngoc Thao

**Version:**

1.0

**Date:**

3<sup>rd</sup> December 2022



# TABLE OF CONTENTS

DOCUMENT VERSION CONTROL .....	2
1. DOCUMENT INFORMATION .....	2
2. DOCUMENT SIGN-OFF .....	2
3. DOCUMENT VERSIONS .....	2
I. DATA INTEGRATION GOALS .....	3
II. DATA INTEGRATION PLAN .....	3
III. DATA INTEGRATION RESULTS .....	4

## DOCUMENT VERSION CONTROL

### 1. DOCUMENT INFORMATION

Document code     **DIP**

Document title     **Data Integration Process**

Version             **1.0**


Authors             **Nhan Nguyen Cao, Trang Ho Ngoc Thao**

Distributed by       **Project VN01 team**

File name           **DIP\_Data Integration Process\_1.0.pdf**

Release definition   **Only released as a finished document**

### 2. DOCUMENT SIGN-OFF

ID	Member	Role	Signature	Timestamp
21142355	Tan Le Tran Ba	Project Manager		5 Dec 2022 23:37

### 3. DOCUMENT VERSIONS

Version	Timestamp	Description	Responsible members
<b>1.0</b>	3 Dec 2022 15:25	Plan and results report for Data Integration Process of the project.	<b>Nhan Nguyen Cao</b> (21142377) <b>Trang Ho Ngoc Thao</b> (21142358)

## I. DATA INTEGRATION GOALS

The purpose of Data Integration Process within the project is to convert the collected data (stored in Google Sheets during the collection phase), format the data and integrated them into ready-to-use format inside MongoDB database for the project. There are two main goals included in the Data Integration Process of the project:

- Insert the data into processing code, and deployed processed data to corresponding documents inside MongoDB database.
- Format and clean data for easy implementation into the system during the development of the features in Development phase.

## II. DATA INTEGRATION PLAN

We defined a pipeline for collecting data for the project, which involves of the following steps:

- Collect the divided items (acupuncture points and meridians), covering the selected fields for collecting (refer to the Research Plan document for more details).
- Input the collected data into Google Sheets on team's Google Drive workspace in appropriate format.

Data Integration is the next step when the collection has been done, to convert the data into ready-to-use states and store inside the Database server of the project. After basic technical research, we selected the following tools and platforms to be used for the Data Integration process:

- Python programming language was selected to be used for Data Integration process. In details, we would develop some notebooks using Jupyter Notebook to execute the integration flow.
- To communicate with the MongoDB database server from Jupyter Notebook, we used the library PyMongo, which is a Python distribution containing tools for working with MongoDB.

A total of three Jupyter Notebooks were developed to handle the whole Integration process, corresponding to the three spreadsheets we used for storing the collected raw data: one for the 60 important acupuncture points, one for the remaining 302 secondary acupuncture points and one for the 14 meridians.

The flow for each Jupyter Notebook is as follows:

1. Insert the raw data from Spreadsheet into the Notebook and store in Dataframe
2. Perform some Data Cleaning steps, and normalize the fields (especially the points and meridians names + codes, used as the identities).
3. Insert the processed data into MongoDB database as documents, to be used for the features.

### III. DATA INTEGRATION RESULTS

Since each team member was assigned two sheets within each file to collect the data in both languages. The first step during Data Integration was to insert all data from all sheets and concatenate them into two collections, one for Vietnamese and one for English. An example of data after collected and added into DataFrame is as follows:

	Acupuncture point code	Acupuncture point name	Acupuncture point description	Acupuncture point location	Acupuncture point functionality	Acupuncture point triggering method	Caution
0	St-25	Thiên xu	Từ huyệt Thần khuyết (giữa rốn) đo ra 2 thốn	Dưới huyệt là gân cơ chéo to, cơ thẳng to, mạc...	1. Viêm phúc mạc\n2. Viêm dạ dày cấp mãn tính\...	Châm thẳng, sâu 1-2 thốn. Cứu 5-7 lửa. Ôn cứu ...	NaN
1	LU-1	Trung phủ	Bờ dưới xương đòn gập, ngang với cơ ngực to. ...	Dưới huyệt là rãnh cơ ngực to, cơ ngực é, cơ r...	1. Viêm khí quản, viêm phổi\n2. Lao phổi\n3. H...	Châm xiên, hướng mũi kim ra ngoài lên trên, sâ...	NaN
2	Du-14	Đại chủy	Khi điểm huyệt ngồi ngay hơi cúi đầu xuống một...	Dưới huyệt là cơ gân thang, cơ gân trâm, gân c...	1. Cứng gáy, vẹo cổ\n2. Nhiệt cấp tính, sốt ca...	Châm thẳng, hơi xiên lên trên 1 - 1,5 thốn, tạ...	Trong trường hợp tiết ứ dịch phế quản khi về ...
3	BI-13	Phế du	Huyệt là nơi gặp nhau của đường thẳng đứng ngo...	Dưới huyệt là cơ thang, cơ trâm, cơ răng cưa b...	1. Viêm khí quản\n2. Suyễn\n3. Viêm phổi, lao ...	Châm thẳng, hơi xiên về phía cột sống, sau 0,5...	Bên dưới là phổi, không châm sâu quá
4	BI-14	Quyết âm du	Huyệt là nơi gặp nhau của đường thẳng đứng các...	Dưới huyệt là cơ thang, cơ trâm, cơ răng cơ bé...	1. Thấp tim\n2. Thần kinh suy nhược\n3. Đau th...	Châm thẳng, hơi xiên xuống đốt sống lưng, sâu ...	NaN

For Preprocessing, we applied some basic steps to handle some following requirements:

- Map the item code and name with the one defined in Wikipedia page for List of acupuncture points (agreed to be used as the base reference for collecting data).
- Split the functionality (for acupuncture points) and list of involved acupuncture points (for meridians), to store this information as an array inside the document in MongoDB.
- Remove some null values and exclude them from the list of keys for each document while formatting for the Database.
- Fix some typo in the values, such as removing duplicated empty spaces or trailing / leading empty spaces or unused characters.
- Convert each row inside the DataFrame into a JSON object, which is compatible to be added into MongoDB database.

The following image illustrates one item that has been preprocessed successfully and is in the final state, ready to be integrated into the Database.

vi[2]

```
{'code': 'GB-20',
 'name': 'Phong trì',
 'description': 'Giao hội huyết của 2 kinh thủ túc Thiếu Dương và mạch Dương duy\nXác định đáy hộp sọ, cơ thang và cơ ức-đòn-chũm. Huyết nằm ở chỗ hõm do bờ trong cơ ức-đòn-chũm và bờ ngoài cơ thang bám vào đáy hộp sọ tạo nên.',
 'anatomy': 'Dưới huyết là góc tạo nên bởi cơ thang và cơ ức-đòn-chũm, đáy là cơ gối đầu và cơ đầu dài. Dưới là đáy hộp sọ. Thần kinh vận động cơ là nhánh của dây cổ 2, nhánh của dây thần kinh chẩm lớn, nhánh của dây thần kinh dưới chẩm. Da vùng huyết chi phối bởi tiết đoạn thần kinh C2.',
 'functionalities': ['Đau đầu',
 'Cứng cổ, cứng gáy',
 'Cảm mạo',
 'Hoa mắt, bệnh mắt',
 'Viêm mũi',
 'Ù tai',
 'Huyết áp cao',
 'Động kinh',
 'Liệt nửa người',
 'Bệnh ở não',
 'Trúng gió'],
 'technique': 'Châm thẳng, ngang với trái tai hơi hướng xuống dưới, khi châm huyết này thì hướng mũi kim qua mắt bên kia, sâu 1 – 1.5 thốn, tại chỗ có cảm giác căng tức có khi giật tới đỉnh đầu, vùng xương vành tai, trước trán hoặc lan ra hố mắt. Châm xiên hướng Phong trì bên kia, sâu 2 – 3 thốn, tại chỗ có cảm giác căng tức, có khi lan ra vùng cổ. Cứ 3 – 7 lửa. Ôn cứu 5 – 10 phút',
 'caution': 'Dưới là hành tủy, không châm sâu'}
```

While the Preprocessing step has been done, the items were integrated into the Database using insert function of PyMongo library. The finished items of acupuncture points and meridians after successfully integrated into MongoDB database are as follows:

```
_id: ObjectId('63d546116d5a116a5a5526e7')
code: "TE-17"
name: "Yifeng"
description: "Posterior to the inferior border of the ear-lobe, in the depression an..."
anatomy: "The needle passes through the skin, subcutis, glandula parotis and rea..."
functionalities: Array
technique: "Let the patient open the mouth and puncture obliquely 1 – 2 cun, directi..."
updatedAt: 2023-03-22T17:23:14.135+00:00
```

```
_id: ObjectId('63d546116d5a116a5a5526e8')
code: "LI-20"
name: "Yingxiang"
description: "Between the nasolabial groove and the midpoint of the lateral border o..."
anatomy: "The needle passes through the skin, subcutis to the musculus quadratus..."
functionalities: Array
technique: "Perpendicularly 0.1 – 0.3 cun, or horizontally 0.5 – 0.8 cun in medial..."
```

```
_id: ObjectId('63d546116d5a116a5a5526e9')
code: "ST-4"
name: "Dicang"
description: "0.4 cun lateral to the mouth angle"
anatomy: "The needle passes through the skin, subcutis, musculus orbicularis ori..."
functionalities: Array
technique: "1.5 – 2 cun horizontally toward point Jaiche. When treating facial par..."
```

In case typographical errors are identified during the implementation of the features or during Quality Assurance step, direct correction would be made to the corresponding document in Database server.