**Alfonso Garcia Miguel**

International Committee of the Red Cross

alfonsocasanueva@gmail.com

**Data Science Project**

**Collaborative work with SharePoint**

**Conceptual Design Report**

**11 October 2020**

# Abstract

This study aims at assessing the level of collaborative work of an institution within its SharePoint database. A sample covering representative services of the institution is analysed in an attempt to describe the whole population and answer this assessment. To this end, the data sample is cleaned and prepared for the analysis, divided in different dataframes and tests are conducted using Python code and Jupyter notebooks. In particular, the following aspects are described: typology of documents, the rhythm of document creation, versioning of documents and service particularities. In order to answer the question if all services work in a similar way, a Mann-Whitney test is run within different samples.

The study concludes that the database is heavily used as a publication platform, but collaboration in it is very poor. Equally, it concludes that the three kind of populations (Operational, Administrative and Political) are sufficiently different to conclude that their way of working is different.

# Table of Contents

[Abstract 0](#_Toc53222346)

[Table of Contents 1](#_Toc53222347)

[1 Project Objectives 2](#_Toc53222348)

[2 Methods 2](#_Toc53222349)

[3 Data 3](#_Toc53222350)

[4 Metadata 5](#_Toc53222351)

[5 Data Quality 5](#_Toc53222352)

[6 Data Flow 6](#_Toc53222353)

[7 Data Model 7](#_Toc53222354)

[8 Risks 8](#_Toc53222355)

[9 Preliminary Studies 8](#_Toc53222356)

[10 Conclusions 9](#_Toc53222357)

[Acknowledgements 10](#_Toc53222358)

# 

# 1 Project Objectives

Over the last six years, SharePoint has been implemented as the document repository for my company. The intention is not only to have a single repository to store documents, but to promote collaborative work among colleagues. The usage of SharePoint was strongly encouraged, but not imposed: Staff had other tools to create documents.

The success of SharePoint implementation was never assessed. It was always assumed that its implementation was successful due to the overwhelming number of documents in the repository (2 million files).

This project aims at assessing if the original objectives of SharePoint deployment have been achieved, namely if staff is working in a collaborative mode (i.e., doing the modifications on the platform itself) or if, on the contrary, staff works in other platforms and use SharePoint only as a publication and storing platform. In order to answer this question, the study will analyze the version history and the size (current and cumulated) of the files in the repository. A document version indicates the number of times the document has been saved in the platform. The more versions a document has, the more collaboration has been achieved in that document. Equally, the cumulated size of a file tells us how this collaboration has been achieved: the bigger the cumulated file size, more additions have been added to the document.

Finally, the study aims at assessing if there are differences regarding the three types of institutional work (Operational, Administrative or Political) and the collaboration degree on these types of work. Typologies of work can be determined by the file types.

# 2 Methods

For the purpose of this study, the following choices have been taken I order to achieve our objectives:

**Software**

**SharePoint:** It is the documentary repository we want to assess. It is where the data we want to analyze resides. SharePoint is delivered by Microsoft;

**Sharegate**: It is the technology needed to extract the data from the SharePoint database. It is provided by Microsoft together with SharePoint, meaning that in the data extraction we are limited to the functionalities provided by this tool;

**Anaconda** **and** **Jupyter** **notebooks**: It is our choice to code our work using python and all related modules

**Hardware**

The usage of a personal computer is enough to achieve the objectives of the project. The heaviest file with row data amounts to 19.5 MB, which is easy computable with my personal laptop.

In addition, a GitHub repository will be used to store copies of the cleaned data files, Jupyter notebooks and the poster. This GitHub copies serve two purposes: sharing the project (and the code) with the data science community and acting as a security back-up in case my personal computer crashes or gets lost.

**Methodology:**

To acquire proper understanding of our data, we will represent it in different ways to globally assess collaboration and how staff work with the platform. We will use histograms by service and file type to assess the degree of collaboration (determining distributions). Scatter plots will show any possible relations between versions and file sizes

Statistically, we will use regression between the cumulated file sizes and version to see if the subsequent versions of a document add content (cumulated file size increases significantly) or if they just correct existing content (cumulated file size does not increases significantly)

In addition, and with the visualization of our data, we will chose an hypothesis testing to obtain a solid response to our question.

# 3 Data

The whole data population in the platform amounts to above 2.000.000 documents, and each day hundreds of documents are added to it. For this project, we will select a sample of all existing data. Although it would be theoretically possible to extract the totality of existing documents in the database to get not a sample, but the whole population, there are two main reasons against this approach:

**Population-wise**: the extraction will indeed reflect the whole population… in a given moment. The database is active and currently in use, new documents are added where others are possibly deleted. This means that minutes after we have extracted the whole population, our data will turn back to just a (big) sample.

**Resources-wise**: it will not be advisable to deal with a volume of more than 2 million observations when we can achieve the same results with a representative sample.

In order to be representative, the sample selected extracts the information from two operational services (Protection and Assistance) two administrative services (Finance and Fundraising) and a governance body (Presidency). (*See Figure 1: Number of documents by service*)

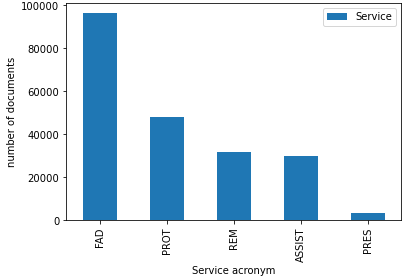


Figure 1: Number of documents by service

Across all data and information that exists in the SharePoint repository of ICRC, the only elements that interest us to answer the question about collaboration are:

**Version number** (low version numbers indicate little collaboration).

**Service** (do some services collaborate more than others?)

**File type** (is there more collaboration on word? On PowerPoint?) (*See Figure 2: Number of documents by file*)

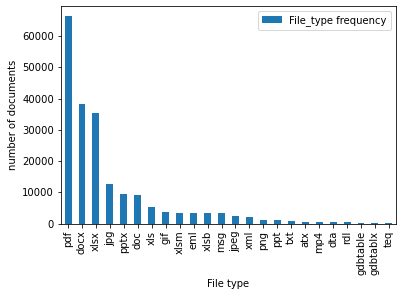


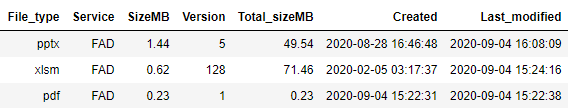
Figure 2: Number of documents by file

**File size** (is there more collaboration on heavy files?)

**Dates of creation** and of **last modification**: (Is there a long time span in which documents stay under edition?).

See an extract of the cleaned data in the table below (*Table 1: extract from clean data*)

Table 1: Extract from clean data



For the purpose of our study, we will consider that true **collaboration starts from version 3**: Creation of the document is version 1, addition of content is version 2, and corrections is version 3. (*See Figure 3: frequency of version number*)

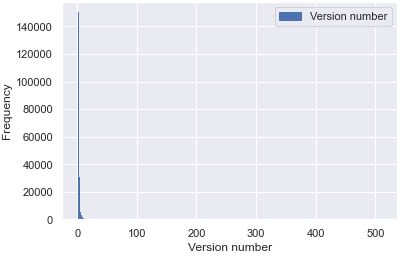


Figure 3: Frequency of version number 1

**Security concerns:**

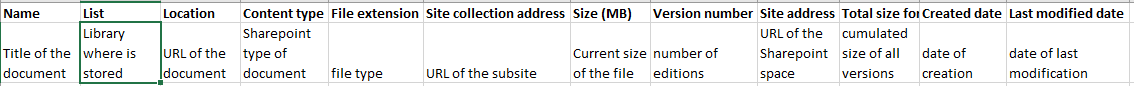
The SharePoint document repository contains documents with sensitive information. With the Sharegate tool, only the title of the document is extracted, not its content. However, even the title can contain sensitive information.

For the purpose of our study, the title does not bring any added value. Therefore, during the cleaning process, this field is removed. The security copies do not contain this field either.

# 4 Metadata

A description of the data is stored together with the extracted files (*See Table 2: Metadata*):

Table 2: Metadata



Sharegate extraction was carried out on the 04/09/2020, on the SharePoint subsites of the following services: PROT, ASSIST, FAD, REM and PRES. The extraction covered the whole database of those services until that date (no time filters were applied.)

To reproduce exactly the same exercise, it is important to use the same subsites and the same dates (first date until 04/09/2020). However, it will be more interesting to run the exercise on a different subset, or on dates from 05/09/2020 to compare the sample with other samples, or to see evolution over time.

# 5 Data Quality

The data sample is extracted from an informatic database, with binary information. Therefore,

Our data is a perfectly accurate picture of the database situation on a given moment. The whole sample is also extracted (ShareGate either produce a complete extraction or prompts an error message) so we can be sure that there is no observations missing.

However, for the purpose of our study, the extraction needs some quality working, namely:

-**Service** can be deducted from the URL of the SharePoint subsites, but it is not a data extracted as such.

-**Dates** are precise to the second. So much precision is not needed.

-**Time difference** between Date of Creation and Date of Last Modification is not provided: Needs to be calculated

-**Titles** of the columns are long, with spaces, and words that can be confused with Python code (‘for’, for instance): A renaming of columns is needed

-**Version** **number** shows two kind of versions: Major Version (published, accessible to all) and Minor Version (not published, accessible to the service only) The format is a floating number where the integer is the major version and the decimal is the minor version. For instance, a version 1.2 indicates major version 1 and minor version 2. A document 0.3 indicates that the document has no major version, but it is on its third minor version. In both cases, the document has been saved three times, and this is the information useful for our study: How many times has a document been saved (modified), regardless of the kind of version.

**Data requirements:**

Cleaned data must fulfill the following requirements:

* Version (after adding major and minor versions) can take any value from 1 to infinite.
* Date of creation must be inferior or equal to date of last modification.
* Columns must be of the following type (*See figure 4: Column types*):

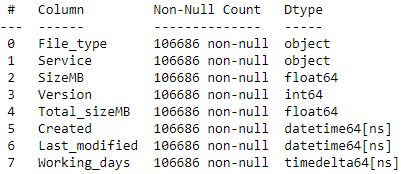


Figure 4: Column types

# 6 Data Flow

Explain with a figure and words how the data flow of your project will be, from the data source to the final plots and numbers and.

SharePoint offers a native import function that list selected metadata (the columns) for each file (the rows, our occurrences) in an excel file. This excel dataset is then uploaded in the Jupyter notebooks to be processed. After the cleaning, a back-up file is stored for security reasons.

The data flow is described in the Figure 5: Data flow

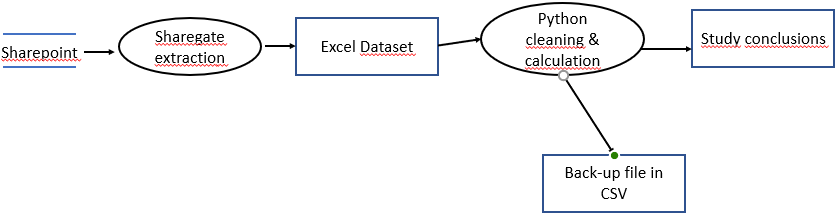


Figure 5: Data flow

# 7 Data Model

Below, an explanation of the Conceptual, Logical and physical models.

**Conceptual Model:**

The data represents the files contained in the SharePoint database, since its creation to the day of extraction. There is no relationship among the occurrences (all are independent from each other) and there are also no relationship among groups of occurrences (for instance, values of files in a service do not affect values of other service)

Values for each row (file) are not repeatable (a document cannot have two versions, two dates of creation, or two services…)

**Logical Model:**

**Datasets**: The cleaning is carried out using the excel dataset as a base. Some plotting is also extracted from this dataset.

**Dataframes**: There are several dataframes:

A general one, with the information from all files

A cleaned one, with only the data interesting for the analysis

Subsets of data, each in its own data frame (one by service, one by file type, one with documents on version 1 to 15)

**Data Structures**:

We used different python modules (Pandas Dataframes, Matplotlib, Numpy, Scipy and Seaborn for some of the plotting)

The original data contains 12 columns and almost 210.000 rows. (Including non-collaborative file types)

The cleaned data contains 8 columns and more than 106.000 rows.

**Physical Model:**

**Database**: SharePoint instance installed locally (Enterprise servers in Geneva. Mirror copy of the servers in Satigny.)

**Dataset**: Extraction of files in excel format: *DocumentData\_ALL.xlsx*. Size 19.5 MB

**Back-up copy**: Cleaned dataset in CSV format: *CollabCleanSPdata.csv*. Size 6.6 MB

Both files copied in the GitHub repository and in my own laptop.

For the **processing**, my laptop is used, as it has enough capacity (32MB memory RAM, intel processor, Windows 10 SO)

# 8 Risks

From a technical point of view, few things can go wrong. As indicated above, we are analyzing a document repository (a data base), which is made of binary, digital information. The database reflects the exact picture of the documents contained in it. There is no risk of external influences or wrong measurements. The only possible concern is that the database does not truly reflect the actual work of staff, but this is exactly what this study is about: *Do people really work in the database or are they just dropping the finalized documents there?*

There are, however, a technical risk: The data we are work with is lost, corrupted or inaccessible. To avoid this situation, once the data is cleaned, it is saved in different copies at different locations.

# 9 Preliminary Studies

This is the first study of its kind. It will certainly serve as a basis to monitor future evolution of the platform, and as a benchmark.

Some interesting findings can be indicated already here:

The distribution of version numbers seems to reflect a *poisson* distribution (to be checked). It is clear that the version numbers of the documents is really low. (*See figure 3: Frequency of version number*)

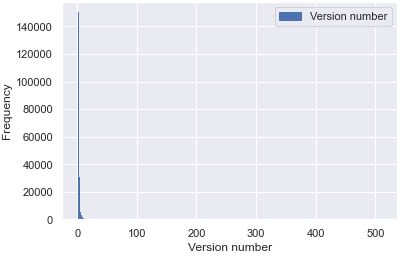


Figure 3: Frequency of version number 2

Having said this, we can guess a different pattern depending on the service the documents belong to. At a later stage we will conduct a Mann-Whitney test to see if we can truly consider our data as coming from different samples or not. In other words, can we consider that services work in a significant different way (they belong to different samples) or not. (*See figure 6: Version distribution by service type*)

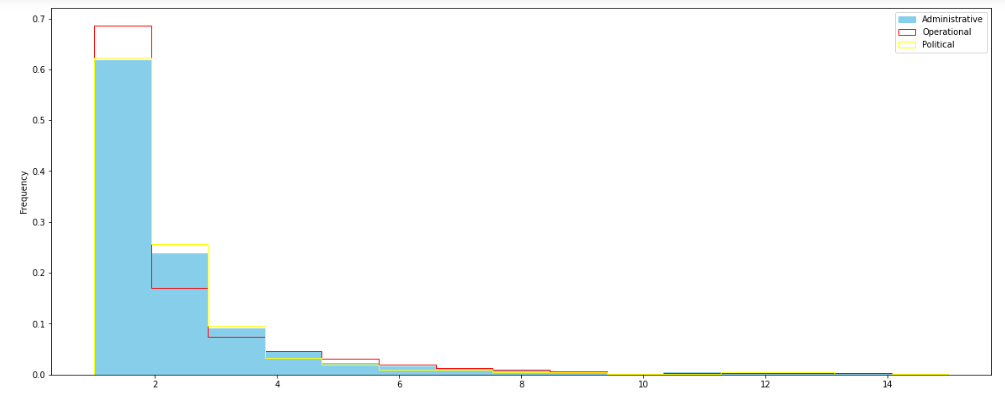


Figure 6: Version by service type

# 10 Conclusions

This is the first study done on the SharePoint database, and in the light of the results, it is fair to conclude the following:

* The database is used generally by all services. The number of documents created in it is significant and relatively stable across time.
* Collaboration is very poor. Although there are some documents updated several times, the vast majority is never updated, or only once. 21.5% of documents reach version 3 or above. (Remember that in chapter 3 we considered that we can start talking about collaboration from version 3). Only a fifth of the documents is really worked in the platform.
* Use for dissemination (publication) is quite high. There is an overall big volume of documents within the database and a significant volume of images and pdf documents.

# Acknowledgements

I would like to acknowledge Goran Stojanovic, for having extracted the data for this study, and my classmate Lionel Perret for the fruitful exchange of ideas.

Most of all, I would like to thank my work colleagues for feeding this database with plenty of documents full of valuable knowledge for the institution.

Finally, I would like to thank the anonymous contributors to “stack-overflow” and similar webpages for the useful information and advice obtained from them.