**DICOM Metadata De-identification Manual**

**INTROUCTION**

The DICOM metadata de-identification process in CHORUS is performed in two steps:

**Step 1: De-identification Using Pydicom Script**

* PatientID and AccessionNumber in the DICOM metadata are replaced by person\_id and image\_occurrence\_id from the OMOP table.
* Selected date tags (Table 1) are shifted by a predefined number of days (specific to each PatientID).

Table 1. Date tags to be shifted.

|  |  |  |
| --- | --- | --- |
| **Group** | **Element** | **Tag\_name** |
| 0008 | 0012 | InstanceCreationDate |
| 0008 | 0020 | StudyDate |
| 0008 | 0021 | SeriesDate |
| 0008 | 0022 | AcquisitionDate |
| 0008 | 0023 | ContentDate |
| 0008 | 002a | AcquisitionDatetime |
| 0010 | 0030 | PatientBirthDate |
| 0018 | 1012 | DateOfSecondaryCapture |
| 0018 | 1078 | Radiopharmaceutical Start DateTime |
| 0018 | 1079 | Radiopharmaceutical Stop DateTime |
| 0018 | 1200 | DateOfLastCalibration |
| 0018 | 700c | DateOfLastDetectorCalibration |
| 0032 | 1000 | ScheduledStudyStartDate |
| 0032 | 1010 | ScheduledStudyStopDate |
| 0032 | 1040 | StudyArrivalDate |
| 0032 | 1050 | StudyCompletionDate |
| 0038 | 0020 | AdmittingDate |
| 0038 | 0030 | DischargeDate |
| 3006 | 0008 | StructureSetDate |

**Step 2: De-identification Using RSNA CTP software**

* Remaining DICOM metadata is de-identified using the RSNA CTP software.
* Pre-identified tags (PatientID, AccessionNumber, selected date tags in Table 1) from Step 1 are preserved.

**INSTRUCTIONS**

**Module 1: Pydicom script**

1. **Prepare lookup tables files: “image\_map.csv” and “personal\_map.csv”**

Templates are located in: “CHORUS\_metadata\_deid\_instruction/pydicom/lookup\_table/”

1. **Run the python code: “pydicom\_deid.py “**
2. Install Pydicom package: pip install pydicom
3. Run the command:

python pydicom\_deid.py --map\_table\_dir <lookup\_table\_dir> --input\_dir <DICOM\_input\_root\_dir> --output\_dir <DICOM\_output\_root\_dir> 2>&1 | tee logs.txt

* <lookup\_table\_dir>: Directory containing the lookup table CSV files.
* <DICOM\_input\_root\_dir> : Root directory of input DICOM files. The script will display de-identification progress for each subfolder within this directory. It is recommended to organize the subfolders by patient for easier tracking.
* <DICOM\_output\_root\_dir>: Root directory where the output DICOM files will be saved. This directory will contain two subfolders:
  + dicom\_processed: Contains successfully de-identified DICOM files.
  + dicom\_unprocessed: Contains DICOM files that could not be de-identified because either the PatientID or AccessionNumber was not found in the lookup tables, or the selected DICOM Date tag is not of type DT (DateTime) or DA (Date).

The output DICOM files will maintain the same relative folder structure as the input directory.

**Module 2: CTP software**

**Reference:**

RSNA CTP Documentation:

[**https://mircwiki.rsna.org/index.php?title=MIRC\_CTP**](https://mircwiki.rsna.org/index.php?title=MIRC_CTP)

1. **Install CTP**

Double-click the “CTP-installer.jar” provided in the folder “CHORUS\_metadata\_deid\_instruction/CTP/” and choose a directory in which to install CTP software.

1. **Configure CTP launcher**

A pre-configured file, “config.xml”, is provided in the CTP folder. This file defines the server setup and processing pipeline, which consists of five stages.

1. **Pipeline overview:**

* ***ArchiveImportService:***

The ArchiveImportService walks the directory tree of a static archive and imports the files it finds. The files are copied from the archive and placed in a separate directory before being passed down the pipeline. When the files have been processed, they are deleted from the directory to which they were copied, but they remain in the archive. To restart the process from the beginning or run a new import, clear the stage's root directory.

For real-time file import, refer to the DirectoryImportService documentation here: [DirectoryImportService - MIRC CTP Wiki](https://mircwiki.rsna.org/index.php?title=MIRC_CTP&_gl=1%2Adunm0m%2A_ga%2AMTU1NzQ5MjAzNC4xNzExOTk1NzQy%2A_ga_EQ32SZ84M3%2AMTcxMTk5NTc0Mi4xLjAuMTcxMTk5NTc0Mi42MC4wLjA.#DirectoryImportService)

* ***DicomFilter***

The DicomFilter, defined in the provided script file “DicomFilter.script”, filters out DICOM files that do not belong to the following modalities:

* + X-ray: computed radiography (CR), digital radiography (DX), mammography (MG)
  + CT
  + MR
* ***IDMap:***

The IDMap constructs map tables for UID elements, AccessionNumber, and PatientID. Note: In this project, only UID mapping is relevant, as PatientID and AccessionNumber are not replaced.

* ***DicomAnonymizer***

The DicomAnonymizer de-identify metadata within DICOM files. The DicomAnonymizer is defined in the provided script file (“DicomAnonymizer\_chorus.script”). In the script, PatientID, AccessionNumber, and selected date tags (Table 1) will remain unchanged.

* ***DirectoryStorageService***

The DirectoryStorageService stores the de-identified DICOM files in the following folder structure defined in the provided configuration file:

*Modality / PatientID / AccessionNumber / SeriesDescription*

1. **Configuration procedure:**
2. ***Configuration file (“config.xml”) modification:***

* Modify “maxThreads” or “port” attributes of the server if necessary.
* Modify pipeline “root” to store processing results as an absolute directory.
* Modify ArchiveImportService “treeRoot” as an absolute directory where your input DICOM data is stored. Modify the value of “minAge” if it is not appropriate.

(b) ***Configuration deployment***

* Replace the default “config.xml” under the root directory of the installed CTP software with your modified version.
* Copy the provided script “DicomAnonymizer\_chorus.script” and “DicomFilter.script” into the subfolder “scripts”.

1. **Output folder**
2. ***Output folder structure***

There will be two subfolders in the defined pipeline root:

* ***quarantines:*** Holds isolated DICOM files that could not be processed due to errors or validation failures.
* ***roots:*** Contains all stage files, including:
  + De-identified DICOM files stored by the DirectoryStorageService stage.
  + ID mapping files (\_map.db and \_map.lg) generated by the IDMap stage.

1. ***View IDMap***

The IDMap can be viewed or downloaded as CSV files through the CTP Home Page, as described in the appendix A of this manual.

For headless Linux systems where accessing the CTP Home Page is not possible, a Java class file (ExportIDMap.class) provided in the CTP folder can be used to convert the mapping database files (\_map.db and \_map.lg) to CSV format.

Steps to convert IDMap files:

* *Download the required JAR file:*

wget https://repo1.maven.org/maven2/jdbm/jdbm/1.0/jdbm-1.0.jar

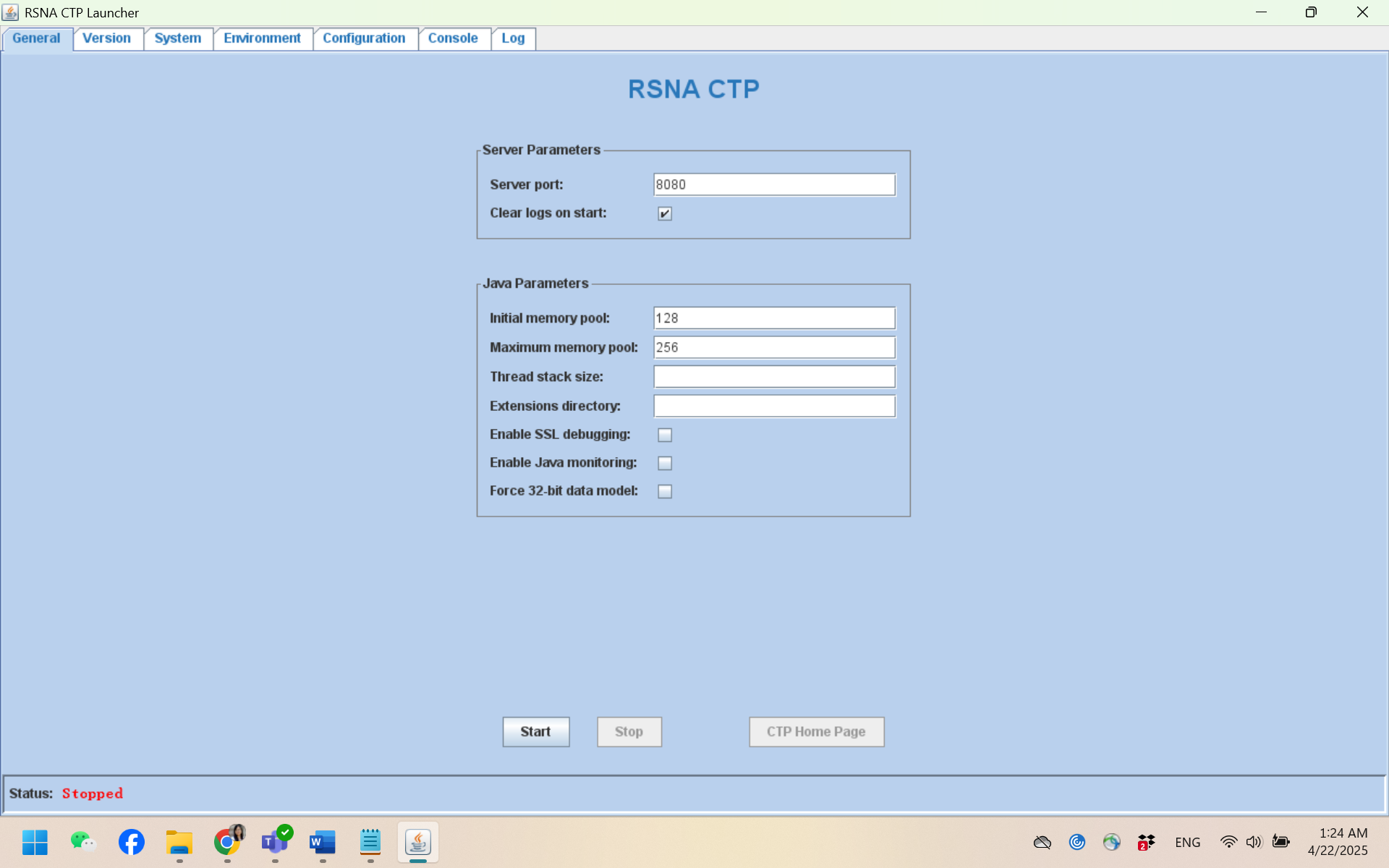
* *Run the code using the following command:*

java -cp .:jdbm-1.0.jar ExportIDMap \_\_map <map\_file\_directory> <output\_directory>

Note: replace "<map\_file\_directory>" and "<output\_directory>" with your actual directories.

1. **Run CTP**

Double-click “Launcher.jar” in the root directory of the software to start the launcher. In the “General” tab, click “Start” to start data processing. To stop data processing, just click “Stop”.



**Testing**

A “test\_sample” folder is provided, containing files and directories intended for reference, testing, or evaluation. The test samples are sourced from the MIMIC dataset, which has already been de-identified. Consequently, the metadata tags included are significantly fewer than those found in typical clinical DICOM images. For additional details, please refer to Appendix B.

**Appendix A: CTP Home Page Introduction**

Click “CTP Home Page” in the “General” tab, it will automatically open the browser and the CTP home page. In this home page, you can check the status of the data processing and ID Map result, and modify the configuration of DicomAnonymizer.

1. **Login in**

If admin has not been set up in the pipeline. The default admin account is:

* Username: admin
* password: password

A screenshot of a computer

Description automatically generated

1. **process status**

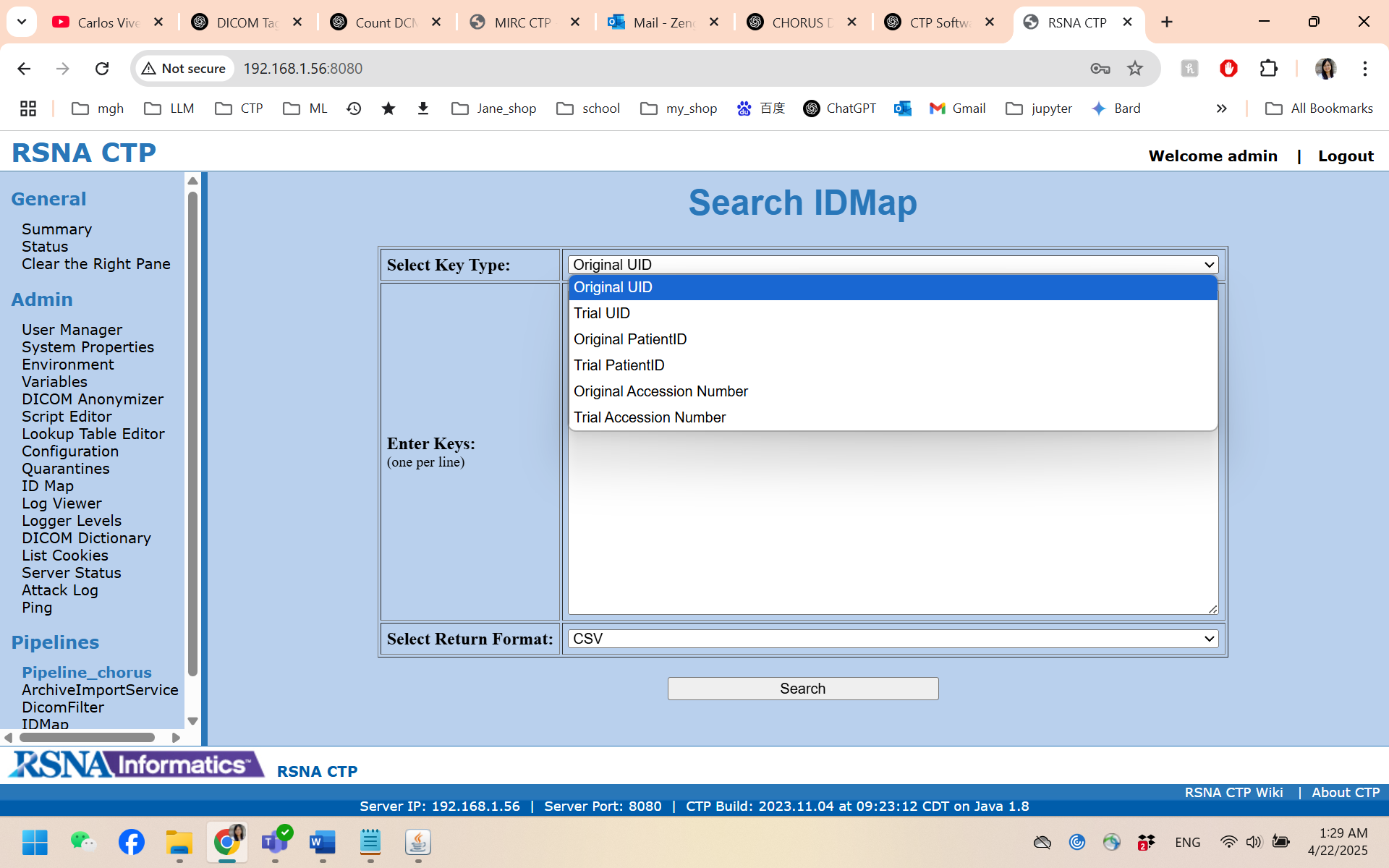
Click “status” on the left panel, you can check the process status.

A screenshot of a computer

Description automatically generated

1. **ID Map**

Click “ID Map” on the left panel to check your interested ID mapping. t also allows us to download a CSV version by selecting “CSV” as the return format:



1. **DicomAnonymizer**

Click “DicomAnonymizer” on the left panel, then click “Edit the Anonymizer Script”, you can check and modify its configuration.

A screenshot of a computer

Description automatically generated

**Appendix B: Testing**

**Test sample structure**

1. **Dicom\_original**

Contains four DICOM files from a single patient, used as original test samples.

1. **Output\_reference**

This folder contains the two-step reference results for the test DICOM images:

* pydicom: Contains the de-identified output from Module 1 (Pydicom script).
* CTPstore: Contains the further de-identified output from Module 2 (CTP software).

1. **UIDmap.csv**

A CSV file generated from the UID mapping files created during the "IDMap" stage in CTP. It records the mappings of UID elements between original and anonymized DICOM files.

1. **compare\_dicom.py**

A Python script used to compare metadata between two DICOM files. It checks for discrepancies in tag and values. If any inconsistencies are found, they will be saved to an Excel file.

**Test & evaluation instructions**

1. **Run Pydicom script**

The lookup tables for the test samples are pre-prepared and available in the “CHORUS\_metadata\_deid\_instruction/pydicom/lookup\_table” directory.

If you are currently in the “CHORUS\_metadata\_deid\_instruction/pydicom” directory and want to save your output to “../test\_sample/output/pydicom” directory, you can run the Python script directly using the following command::  
*python pydicom\_deid.py --map\_table\_dir ./lookup\_table --input\_root\_dir ../test\_sample/dicom\_original --output\_root\_dir ../test\_sample/output/pydicom 2>&1 | tee logs.txt*

1. **Run CTP software**

Modify the “root” attribute in the pipeline element and “treeRoot” attribute in the ArchiveImportService element within the configuration file “config.xml”, then deploy the configuration as illustrated on page 3. Once the de-identification process in step 1 is completed, launch the CTP software and initiate the de-identification process.

1. **Evaluation**

After completing the de-identification process, use the “compare\_dicom.py” script to verify whether the metadata in your de-identified DICOM images matches the provided reference output.

Steps to run the script:

* Download the required python package if they are not installed:

*pip install pandas openpyxl*

* Run the code using the following command:

*python compare\_dicom.py <dicom\_file1\_path> <dicom\_file2\_path> <output\_path>*

* + <output\_path> : Path where the Excel file with any detected differences will be saved.