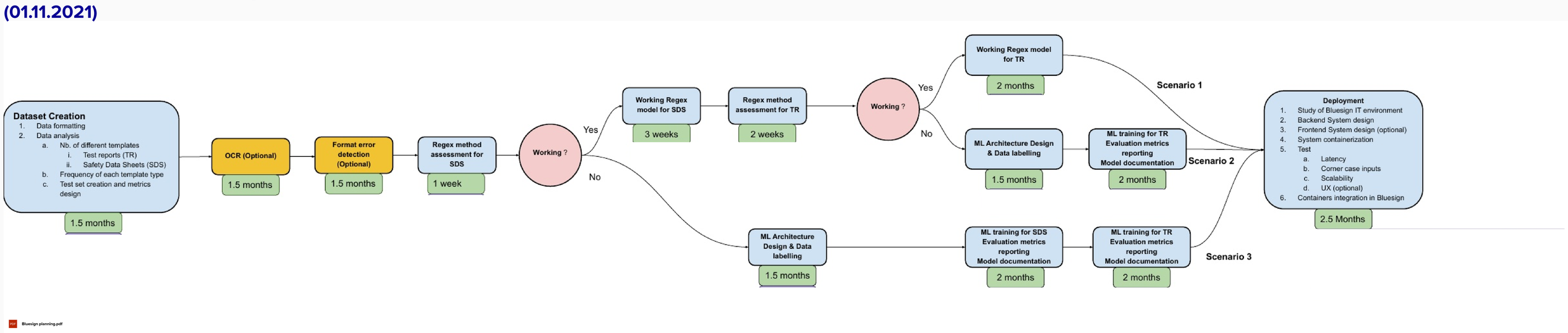
Updated version of the bluesign® scenaris (01.11.2021)



Bluesign planning.pdf

SCENARIO 1 :

Dataset Creation

1.5 months

Data collection
Data Analysis

OCR (optional)

1.5 months

“OCR” Optical Character Recognition is a powerful technological process that converts a paper document, PDF or even a photo into an editable digital document. Automated data extraction

The program behind this technology starts by analyzing the structure of the image file, it tries to separate the different structures that compose the image: tables, texts, images, etc.

Then each line is converted into words and then into characters. Once the characters have been defined, the OCR begins to compare them with a group of images already processed by the OCR or predefined which allow it to propose hypotheses on the meaning of these characters.

The OCR uses these assumptions to analyze the curvatures of lines in words and words in characters. After reviewing all the assumptions, the program ultimately proposes content that could be consistent with the initial image.

Format error detection (optional)

1.5 months

Format errors, in the best-case scenario, can break an automated data processing pipeline. In the worst case, they introduce logical errors in downstream analytical tasks that are difficult to debug.

Example : international different way to write dates.

In our case, the documents are received in different formats.

Regex method assessment for SDS

1 week

=> Focus on the Safety Data Sheet

Regular Expression – better known as Regex – is pattern matching. They are used to match well-defined patterns (= matching tools identify words and phrases with common patterns). It is not an AI solutions using Natural Language Processing (NLP) and Machine Learning (ML) that can conduct an intelligent analysis of unstructured text. Indeed, NLP understands the language as it allows computers to undertake linguistic analysis and in practical terms “read” a document just like a human. In our situation, we don't need to understand the text. We need to report the content of document which are always following the same pattern.

Apply the Regex model on SDS

3 weeks

Apply the SDS regex model on all SDS documents.

Regex method assessment for TD

2 weeks

=> Focus on the Test Report

Regular Expression – better known as Regex – is pattern matching. They are used to match well-defined patterns (= matching tools identify words and phrases with common patterns). It is not an AI solutions using Natural Language Processing (NLP) and Machine Learning (ML) that can conduct an intelligent analysis of unstructured text. Indeed, NLP understands the language as it allows computers to undertake linguistic analysis and in practical terms “read” a document just like a human. In our situation, we don't need to understand the text. We need to report the content of document which are always following the same pattern.

Apply the Regex model on TR

2 months

Apply the Regex model on all Test Report (TR) documents.

Deployment

2.5 months

Models deployment in your IT environment.

SCENARIO 2 :

Dataset Creation

1.5 months

Data collection
Data Analysis

OCR (optional)

1.5 months

“OCR” Optical Character Recognition is a powerful technological process that converts a paper document, PDF or even a photo into an editable digital document. Automated data extraction

The program behind this technology starts by analyzing the structure of the image file, it tries to separate the different structures that compose the image: tables, texts, images, etc.

Then each line is converted into words and then into characters. Once the characters have been defined, the OCR begins to compare them with a group of images already processed by the OCR or predefined which allow it to propose hypotheses on the meaning of these characters.

The OCR uses these assumptions to analyze the curvatures of lines in words and words in characters. After reviewing all the assumptions, the program ultimately proposes content that could be consistent with the initial image.

Format error detection (optional)

1.5 months

Format errors, in the best-case scenario, can break an automated data processing pipeline. In the worst case, they introduce logical errors in downstream analytical tasks that are difficult to debug.

Example : international different way to write dates.

In our case, the documents are received in different formats.

Regex method assessment for SDS

1 week

=> Focus on the Safety Data Sheet

Regular Expression – better known as Regex – is pattern matching. They are used to match well-defined patterns (= matching tools identify words and phrases with common patterns). It is not an AI solutions using Natural Language Processing (NLP) and Machine Learning (ML) that can conduct an intelligent analysis of unstructured text. Indeed, NLP understands the language as it allows computers to undertake linguistic analysis and in practical terms “read” a document just like a human. In our situation, we don't need to understand the text. We need to report the content of document which are always following the same pattern.

Apply the Regex model on SDS

3 weeks

Apply the SDS regex model on all SDS documents.

Regex method assessment for TD

2 weeks

=> Focus on the Test Report

Regular Expression – better known as Regex – is pattern matching. They are used to match well-defined patterns (= matching tools identify words and phrases with common patterns). It is not an AI solutions using Natural Language Processing (NLP) and Machine Learning (ML) that can conduct an intelligent analysis of unstructured text. Indeed, NLP understands the language as it allows computers to undertake linguistic analysis and in practical terms “read” a document just like a human. In our situation, we don't need to understand the text. We need to report the content of document which are always following the same pattern.

ML Architecture Design & Data labelling

1.5 months

The data labeling is the key step that will allows us to identify the screened data and provide one or more meaningful and informative labels. This will help the Machine Learning model in its learning process.

The organization of this algorithm will be personalised for the purpose of Bluesign needs and expectations.

ML training for TR Evaluation metrics reporting Model documentation

2 months

This is the final step = combinaison of the raw data analysis, features selection, machine learning model settings. Combined, and applyed to the training set of the data (coming from your archive and current dataset), we will apply our model on the Test Report document.

Deployment

2.5 months

Models deployment in your IT environment.

SCENARIO 3 :

Dataset Creation

1.5 months

Data collection
Data Analysis

OCR (optional)

1.5 months

“OCR” Optical Character Recognition is a powerful technological process that converts a paper document, PDF or even a photo into an editable digital document. Automated data extraction

The program behind this technology starts by analyzing the structure of the image file, it tries to separate the different structures that compose the image: tables, texts, images, etc.

Then each line is converted into words and then into characters. Once the characters have been defined, the OCR begins to compare them with a group of images already processed by the OCR or predefined which allow it to propose hypotheses on the meaning of these characters.

The OCR uses these assumptions to analyze the curvatures of lines in words and words in characters. After reviewing all the assumptions, the program ultimately proposes content that could be consistent with the initial image.

Format error detection (optional)

1.5 months

Format errors, in the best-case scenario, can break an automated data processing pipeline. In the worst case, they introduce logical errors in downstream analytical tasks that are difficult to debug.

Example : international different way to write dates.

In our case, the documents are received in different formats.

Regex method assessment for SDS

1 week

=> Focus on the Safety Data Sheet

Regular Expression – better known as Regex – is pattern matching. They are used to match well-defined patterns (= matching tools identify words and phrases with common patterns). It is not an AI solutions using Natural Language Processing (NLP) and Machine Learning (ML) that can conduct an intelligent analysis of unstructured text. Indeed, NLP understands the language as it allows computers to undertake linguistic analysis and in practical terms “read” a document just like a human. In our situation, we don't need to understand the text. We need to report the content of document which are always following the same pattern.

ML Architecture Design & Data labelling

1.5 months

The data labeling is the key step that will allows us to identify the screened data and provide one or more meaningful and informative labels. This will help the Machine Learning model in its learning process.
The organization of this algorithm will be personalised for the purpose of Bluesign needs and expectations.

ML training for SDS Evaluation metrics reporting Model documentation

2 months

This is the final step = combinaison of the raw data analysis, features selection, machine learning model settings. Combined, and applyed to the training set of the data (coming from your archive and current dataset), we will apply our model on the Safety Data Sheet document.

ML training for TR Evaluation metrics reporting Model documentation

2 months

Same thing than the previous step, but for Test Report instead of Safety Data Sheet.

Deployment

2.5 months

Models deployment in your IT environment.