

ДОДАТОК А

Слайди презентації

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Атестаційна робота магістра

**Дослідження методів обробки даних у документах
формату docx для визначення помилок**

Науковий Керівник:
доц. к.т.н.

Ревенчук І. А.

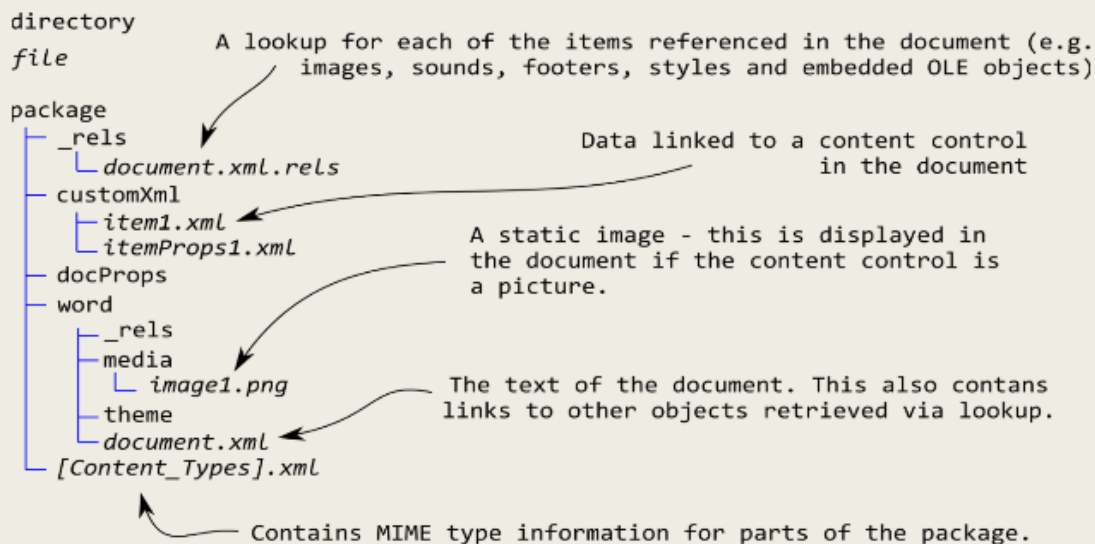
Виконав:
студент групи ШЗм-18-3

Будянський О. О.

Мета роботи

- Дослідження основних методів автоматизації обробки даних у документах формату docx для знаходження помилок форматування
- Дослідження основних елементів структури формату docx
- Розробка алгоритмів валідації форматування документу формату docx
- Розробка програмної системи для автоматизованого нормоконтролю

Структура формату docx



Методи обробки документів docx

COM-
об'єкти



Недоліки:

- залежність від процесу Microsoft word
- низька швидкість
- низька надійність

Об'єкти
.NET.



Недоліки:

- усі існуючі рішення платні
- низька гнучкість

Власна реалізація обробки документів




Недоліки:

- складність розробки


Алгоритм валідації відповідності заголовків змісту сторінок

1 Конвертація



2 Аналіз структури змісту, парсинг та формування списку заголовків до їх сторінок в docx 

```
<w:fldChar w:fldCharType="begin"/>
</w:fldChar>
<w:fldChar w:fldCharType="begin"/>
</w:fldChar>
<w:instrText xml:space="preserve"> TOC \* MERGEFORMAT
</w:instrText>
</w:fldChar>
<w:fldChar w:fldCharType="separate"/>
</w:fldChar>
<w:fldChar w:fldCharType="end"/>
</w:fldChar>
```

3 Пошук вказаних заголовків у основній частині документа docx 

4 Пошук вказаних заголовків на відповідних сторінках у pdf 

Алгоритм валідації кількості контенту

1. Конвертувати docx в pdf



2. Для кожної сторінки

а. Знайти розмір сторінки в pdf



б. Знайти позицію останнього тексту

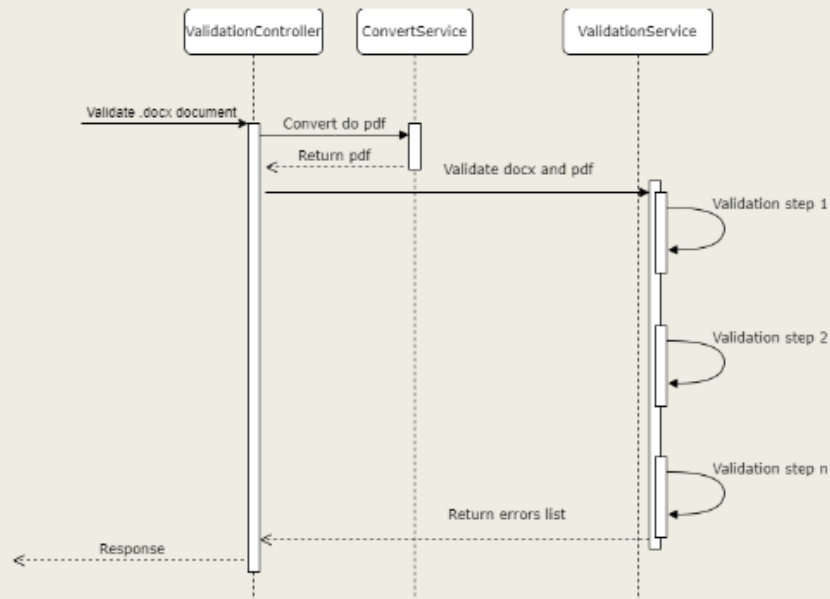


в. Знайти відношення

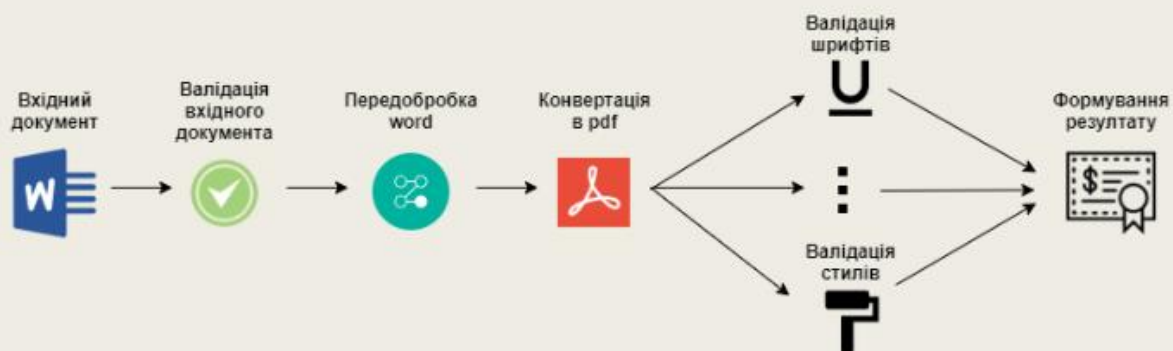
$$700 / 800 * 100 \% = 87.5\%$$

87.5% > 80% - Контенту достатньо

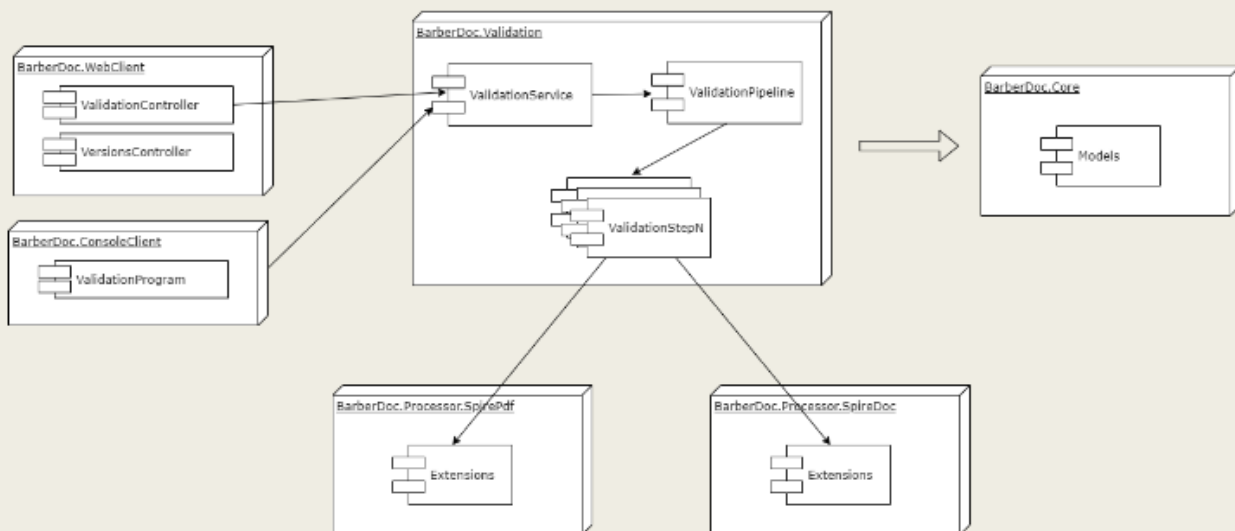
Діаграма послідовності



Паралелізація



Структура проекту



Технології



Приклад коду для знаходження елементів змісту

```

1 reference | Alexey Budynsky, 10 days ago | 1 author, 1 change | 1 work item | 0 exceptions
private IDictionary<string, string> GetTitlePerPageNumber(Document document, IList<ValidationError> validationErrors)
{
    var titleParagraphs = _contentSectionHelper
        .GetSectionItems(document, _validationSettings.SectionTitle)
        .Where(sectionItem =>
            sectionItem.IsParagraph.paragraph
            && !string.IsNullOrEmpty(paragraph.Text)
            && paragraph.StyleName.Contains(TableOfContentStyleNamePrefix))
        .Select(sectionItem => (Paragraph)sectionItem);

    var titlePerPageNumber = titleParagraphs
        .Where(currentParagraph => IsTitleUnique(validationErrors, titleParagraphs, currentParagraph))
        .ToDictionary(
            title => title.Text.Split(TableOfContentSeparator).First(),
            title => title.Text.Split(TableOfContentSeparator).Last());


    return titlePerPageNumber;
}

```

Програмна реалізація


Welcome back to BarberDoc

Step 1: Upload your work How does it work?



Drag and drop MS Word file here
or
[Browse for file](#)

Diploma_Oleksii_Budynsky_v1.docx
158.43 KB



Drag and drop PDF file here
or
[Browse for file](#)

Diploma_Oleksii_Budynsky_v1.pdf
872.12 KB

[Start validation](#)
[Clear results](#)

Step 2: Analyze the results Show validation settings

#	Type	Message	Source
1	Error	Page 5 should not contain page number.	PageNumberingValidationStep
2	Error	Page 3 should not contain page number.	PageNumberingValidationStep
3	Error	Page 1 should not contain page number.	PageNumberingValidationStep

Тестування

Тест сценарій	Кроки	Очікуваний результат	Актуальний результат	Статус
Отримати список помилок форматування файлу диплома.	<ol style="list-style-type: none"> 1. Відкритий сторінку сайту. 2. Взяти валідну атестаційну роботи, та видалити один з заголовків, що вказаний у змісті. 3. Завантажити цей docx файл атестаційної роботи. 4. Завантажити pdf файл цієї атестаційної роботи. 5. Натиснути кнопку «Validate» 	Отримати помилку невідповідності змісту до реальних заголовків.	Отримано помилку невідповідності змісту до реальних заголовків.	Пройдено

Висновки

- Функціоналу одного формату docx недостатньо для повноцінної валідації форматування
- Для більшої ефективності слід поєднувати переваги декількох форматів (docx + pdf)
- Створений програмний продукт може значно економити час та ресурси та є затребуваним у різних сферах

ДОДАТОК Б

Апробація результатів роботи

60 Jubileuszowa Konferencja Studenckich Kół Naukowych ● Pionu Górniczego AGH

MATERIAŁY KONFERENCYJNE

Kraków, 5 grudnia 2019 r.



Główny Partner

SPIS TREŚCI

Sekcja I	Górnictwo, Wiertnictwo, Nafta i Gaz	3
Sekcja II	Geologia	15
Sekcja III	Geoturystyka i Geofizyka	32
Sekcja IV	Geodezja, Kartografia i Geoinformacja	40
Sekcja V	Budownictwo	51
Sekcja VI	Chemia i Ochrona Środowiska	61
Sekcja VII	Wentylacja, Klimatyzacja i Ogrzewnictwo	79
Sekcja VIII	Inżynieria Mechaniczna	94
Sekcja IX	Energetyka	113
Sekcja X	Akustyka	120
Sekcja XI	Informatyka	138
Sekcja XIIa	Inżynieria Produkcji i Jakości	170
Sekcja XIIb	Inżynieria Produkcji i Jakości	188
Sekcja XIII	Zarządzanie i Marketing	206
Sekcja XIV	Przedsiębiorczość Zrównoważona i Innowacyjna .	222
Sekcja XV	Rachunkowość i Finanse	243
Sekcja XVI	Technology, Society and Language	251

Oleksii BUDIANSKYI
Serhii IORDANOV

Koło Naukowe Modelowania w Finansach

Software Engineering Department at Kharkiv
National University of Radio Electronics



Koło naukowe

Modelowanie w finansach

USING ELK STACK AND. NET CORE TECHNOLOGIES IN CASE ANALYSIS OF THE UKRAINIAN VEHICLES OPERATIONS DATASET

Recently Ukraine government has created a portal of open data with different types of data sets related to versatile operations inside the country for all citizens. One of the largest open dataset of this site is a register of vehicles and their owners which contains all information about car operations such as initial registration, registration of imported or re-registration of a vehicle to a new owner. All operations related to vehicles are listed in the dataset. Single data row contains main information about the car (color, registration plate, capacity, brand, etc.), the owner (registration address, individual registration type) and the operation information (name, code).

This paper contains results of format the data to one style, collecting it in one database, analyze and visualization with the most interesting findings. At first to analyze all the data and find any dependencies it's need to formatted them and collected in one database, to be able to compare different records. There were about 11 million data records that should be filtered and aggregated quickly with the ability to display the results of aggregation as charts. After investigation of the available tools and databases Elasticsearch was chosen as the main database.

Elasticsearch is a distributed, RESTful search and analytics engine capable of addressing a growing number of use cases, it lets perform and combine many types of searches – structured, unstructured, geo, metric, etc. So, it is perfect for storing a huge number of data and performing searches. Elasticsearch use of Lucene Core under the hood which provides Java-based indexing and search technology, as well as spellchecking, hit highlighting and advanced analysis/tokenization capabilities. In addition, there is a great visualization tool for Elasticsearch named Kibana, which is an open-source data visualization dashboard for the database. It provides visualization capabilities on top of the content indexed on an Elasticsearch cluster.

In the register, the information was kept in several tar.gz archives which contained csv files with the operation rows. It was not a convenient way to analyze the data and the chosen database was tuned for JSON format. So, the next challenge was converting and transferring the data to the local Elasticsearch cluster.

To perform the operation we have created a C# tool, which parses csv data row, depending on a format, as files have a little bit different ways of naming, then converts data to standard view to be capable of making a comparison, converts the resulting entity to JSON format and sends to Elasticsearch. In order to perform that transformation, we have used CsvHelper and Json.NET (formally known as Newtonsoft JSON). They both are distributed under the licenses which are free for commercial use (MS-PL and Apache 2 for CsvHelper and MIT for Json.NET).

The created solution did the job correctly however the data processing speed was quite slow. The tool was able to process 20-25 data records per second on the local computer, which meant that 11 million data rows would have been processing for more than 150 hours. The most time-consuming operations were parsing the data from csv and converting the data to JSON. As all datafiles were approximately the same size, we decided to split the threads by data source.

The parallel solution worked much better and the throughput was about 100 data rows per second, so all the data were transferred to the Elasticsearch database in around 30-35 hours and became ready to be analyzed. The paper presents interesting results of the Ukrainian Vehicles Operations Dataset analysis with many original illustrations.

Opiekun naukowy:

Associated professor, Volodymyr Kobziev PhD

ДОДАТОК В

Специфікація програмного продукту

Software Requirements Specification

for

BarberDoc Project

Version 1.0 approved

Prepared by

Serhii Iordanov

Oleksii Budianskyi

NURE

04/07/2020

Revision History

Date	Description	Author	Comments
04/03/2020	v.0.1	Serhii Iordanov	Initial draft
04/05/2020	v.0.2	Oleksii Budianskyi	Review + enhancements
04/07/2020	v.1.0	Serhii Iordanov Oleksii Budianskyi	Final Revision

Document Approval

The following Software Requirements Specification has been accepted and approved by the following:

Signature	Printed Name	Title	Date
	Ilona Revenchuk	Product Owner	04/07/2020

1 INTRODUCTION

Nowadays, there are many institutions that work with a constant stream of documents that must meet certain well-defined formatting and layout requirements. Most of these organizations are faced with the problem of document clearance, which is a routine job and takes up a large part of the working hours of specialists. Similar problems can be encountered in almost every field, as document management is an integral part of the modern profession. Solution is based on the problem of automatic documents formatting verification.

Web-service devoted to automate the documents formatting verification process. System receives the input document and formatting rules and returns the list of validation errors to fix.

1.1 Purpose

The SRS document devoted to specify functional and non-functional requirements to cover the business needs in scope of the system described above.

1.2 Scope

The BarberDoc Project scope includes the next parts:

- Implement the library for document validation.
- Implement expandable set of validators.
- Implement an API to work with the library from the Web.
- Implement the UI application for end users.
- Implement a console application to test and develop the library.

1.3 Definitions, Acronyms, and Abbreviations

API – Application Programming Interface

XML – Extensible Markup Language

PDF – Portable Document Format

DFD – Data Flow Diagram

2 GENERAL DESCRIPTION

This section of the SRS describes the general factors that affect 'the product and its requirements.

2.1 Product Perspective

There huge clients' base. The system could be used by any person who works with documentation. We have an ability to enter into a contract with universities and other organizations.

2.2 Product Functions

The system will consist of three main parts: frontend, backend and validation library. In production, the service should provide the following possibilities:

1) Backend

- a. Handles authentication;
- b. Performs preliminary request validation;
- c. Deserializes the document from the request;
- d. Calls validation library;

- e. Serializes validation result;
- f. Performs localization.
- g. Returns validation results as JSON

2) Frontend

- a. Displays validation errors;
- b. Provides inputs for files uploading;
- c. Provides an access to FAQ;
- d. Provides an access to formatting settings.

3) Validation Library

- a. Executes formatting validators concurrently;
- b. Combines validators results into a single list of errors.

2.3 User Characteristics

There are two main stakeholders:

- Students – need to have an ability to preliminary validate the formatting of their works which are in progress or ready to be finally verified.
- Formatting validation experts - need to automatically validate student's works.

2.4 General Constraints

Development limited by the following restrictions:

- development will be conducted only for NURE university;
- the trial license of Spire.Doc and Spire.Pdf is used.

2.5 Assumptions and Dependencies

The system supports only docx and pdf format for input documents and currently configured to work with “ДСТУ 3008-2015” requirements.

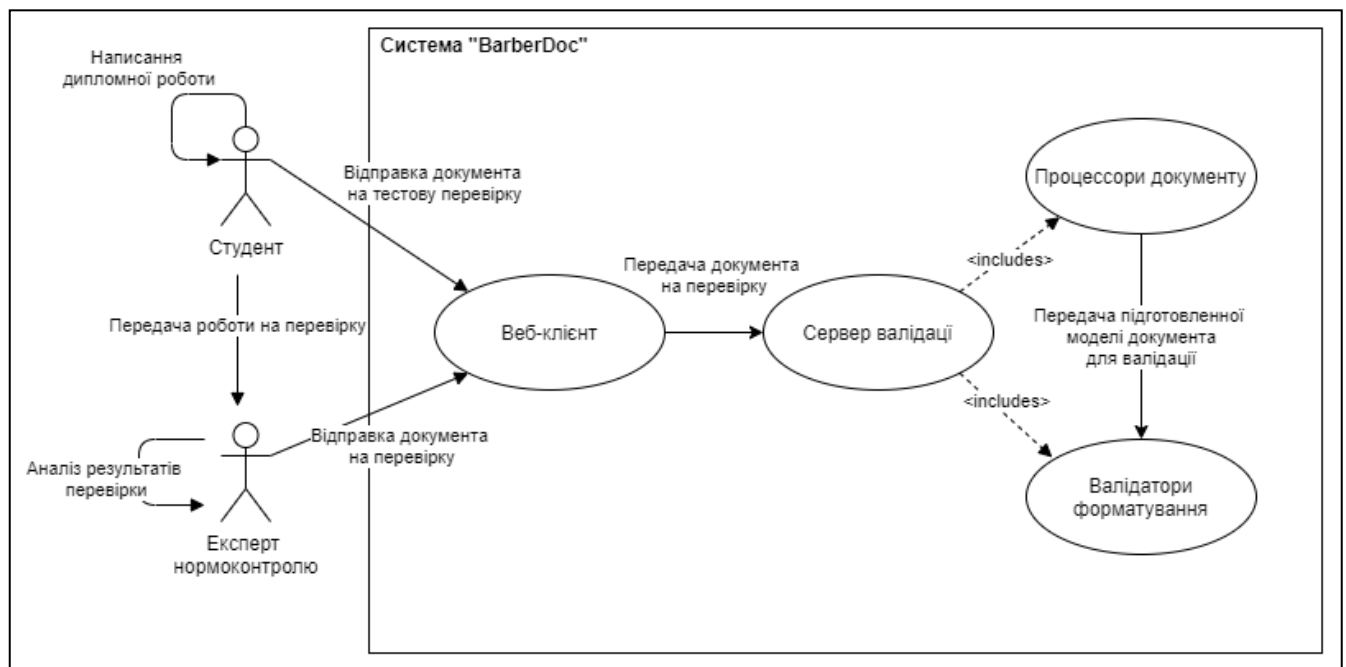
3 SPECIFIC REQUIREMENTS

3.1 User roles

BarberDoc solution addresses the requirements for the following group of users:

- Students
- Formatting validation experts
- System administrator

3.2 Use case scenarios



Picture 1 – Use case diagram

Main actors:

- Student;
- Formatting validation experts.

Pre-conditions

- Student has the prepared work in docx format;
- System administrator had configured the system to correctly validate the formatting;

Main flow

- Students sends the work to the system for preliminary validation;
- System provides validation errors;
- Student fixes the formatting of the work;
- Student sends the work to the formatting validation expert;
- Formatting validation expert validates the work via the system.

3.3 Functional requirements

Document - a file of the .docx format that meets the following requirements.

Sections and subsections are marked with the “Title” tool of MS Office.

Functional requirements:

- a) The system should have several ways of interacting with the user:
 - 1) Through a console application:
 - A Document is submitted to the input;
 - The output is a list of errors.
 - 2) Via the web client:
 - A Document is submitted to the input;
 - The output is a list of errors.
- a) The system should analyze the Document for a violation of the specified formatting rules:
 - 1) The document should not be blank
 - 2) The first page of the domain should be the title page
 - No page numbering;

- There is a hat "Міністерство ...";
 - At the end of the page, the current year is centered.
- 3) There should be a page with the content
- The page contains a Table of Content element;
 - Page numbering matches the real;
 - Items are arranged in order and match the document headers;
 - All items are written in the format of sentences (The first letter is large, then small).
- 4) Before the page with the content there is a page with the title "РЕФЕРАТ / ABSTRACT"
- The first paragraph of the text correctly indicates the number of figures, tables, diagrams, applications, sources
- 5) The indentation throughout the note is five characters or 1.25 cm. Fields:
- left - 2.5 cm.
 - right - 1 cm
 - upper and lower - 2 cm.
- 6) Formatting throughout the document
- Interval of the main text - one and a half;
 - Text font - Times New Roman, black, 14;
 - Throughout the Document, starting from the "VSTUP" section, there should be numbering (in the upper right corner);
 - The page should not end with a title;
 - Using the correct dash;
 - The page cannot begin with a picture.
- 7) Sections:
- Each section starts on a new page;
 - The last page of the section must have at least 10 lines;
 - The section cannot end with a picture, table, diagram;

- Headings of the main sections - in capital letters;
- Headings and subheadings do not end with a period;
- Section title - indent or in the center;
- Section headings - with indentation;
- Two indents after the title or subtitle before the text (normal indentation between the title and subtitle);
- Numbering of headings and subheadings without a dot (not "1.1.", But "1.1");
- Partition numbering - in order of placement.

8) Applications:

- The appendices are at the end of the Document and are numbered A, B, C ...;
- Application formatting is ignored.

9) List of sources

- Numbered;
- References to sources are present in the text (at least one reference to the source), in the format "[1]";
- Arranged in the order of mention in the text.

10) Lists

- Only letters, numbers, or dashes are used.

11) Pictures

- Numbering - end-to-end or section by section;
- In applications, numbering - A.1, A - application number;
- The signature of the figure is "Figure 1.1 - The name of the figure";
- Indentation before the drawing, after the drawing and before the text after the signature;
- The text should contain a link to the picture;
- The figure is in the center.

- 12) Tables:
- Table name - above the table;
 - The name format is “Table 1.1 - Table Name”;
 - The table column names are located in the center of the table cell;
 - Indentation before and after the table.
- 13) The result of the analysis should be a list of errors and warnings about found formatting violations.

4 NON-FUNCTIONAL REQUIREMENTS

4.1 Performance

The system must be interactive and the delays involved must be less than 1 minute. So, in every action-response of the system, there are no immediate delays. In case of opening pages, of popping error messages and saving the settings or sessions there is delay much below 2 seconds. Also when connecting to the server the delay is based editing on the distance of the 2 systems and the configuration between them so there is high probability that there will be or not a successful connection in less than 30 seconds.

4.2 Reliability

As the system provide the right tools for discussion, problem solving it must be made sure that the system is reliable in its operations and for securing the sensitive details.

4.3 Availability

If the internet service gets disrupted while sending information to the server, the information can be send again for verification.

4.4 Security

The main security concern is for users account hence proper login mechanism should be used to avoid hacking. The tablet id registration is way to spam check for increasing the security. Hence, security is provided from unwanted use of recognition software.

4.5 Usability

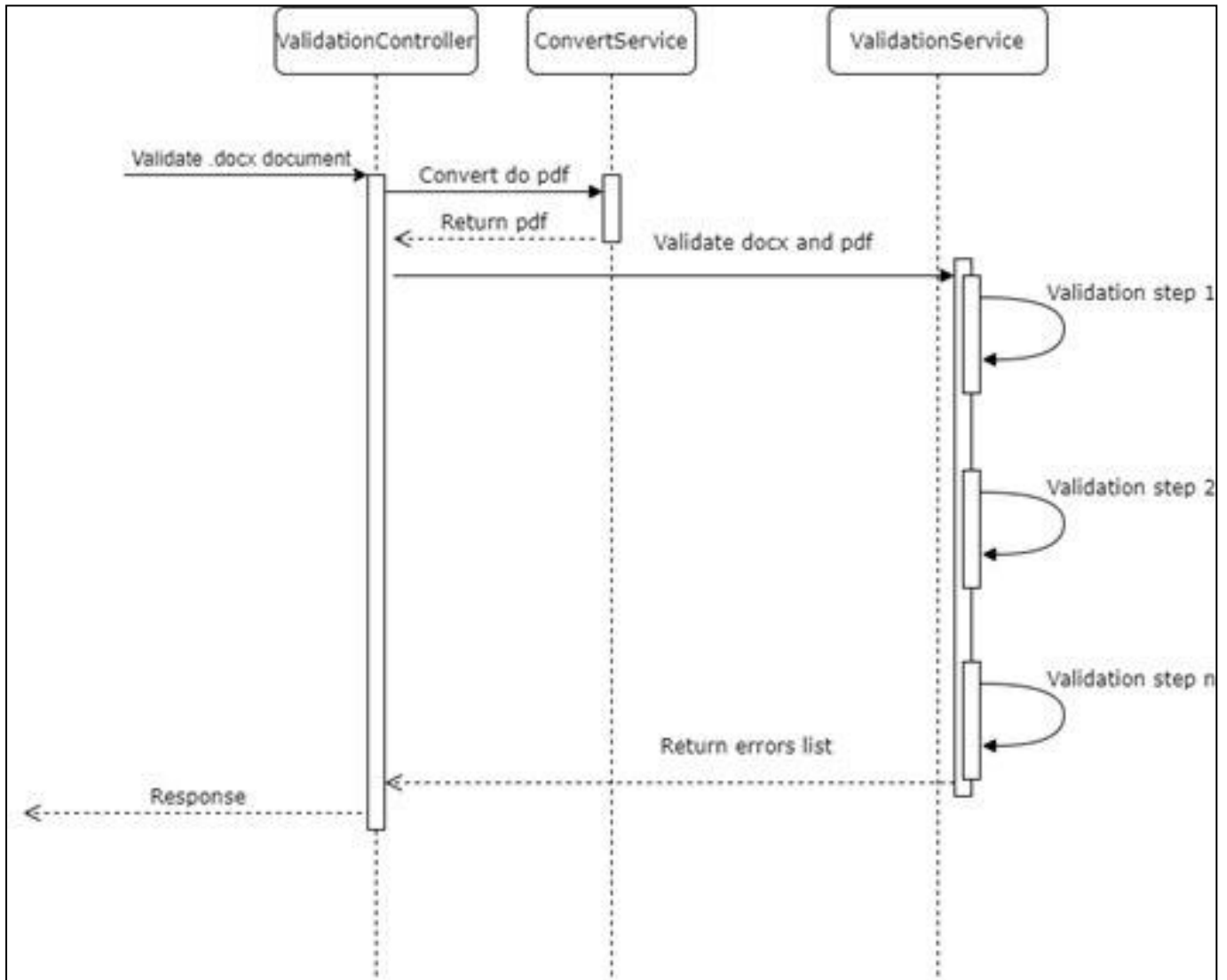
As the system is easy to handle and navigates in the most expected way with no delays. In that case the system program reacts accordingly and transverses quickly between its states.

5.6 Safety

Information transmission should be securely transmitted to server without any changes in information.

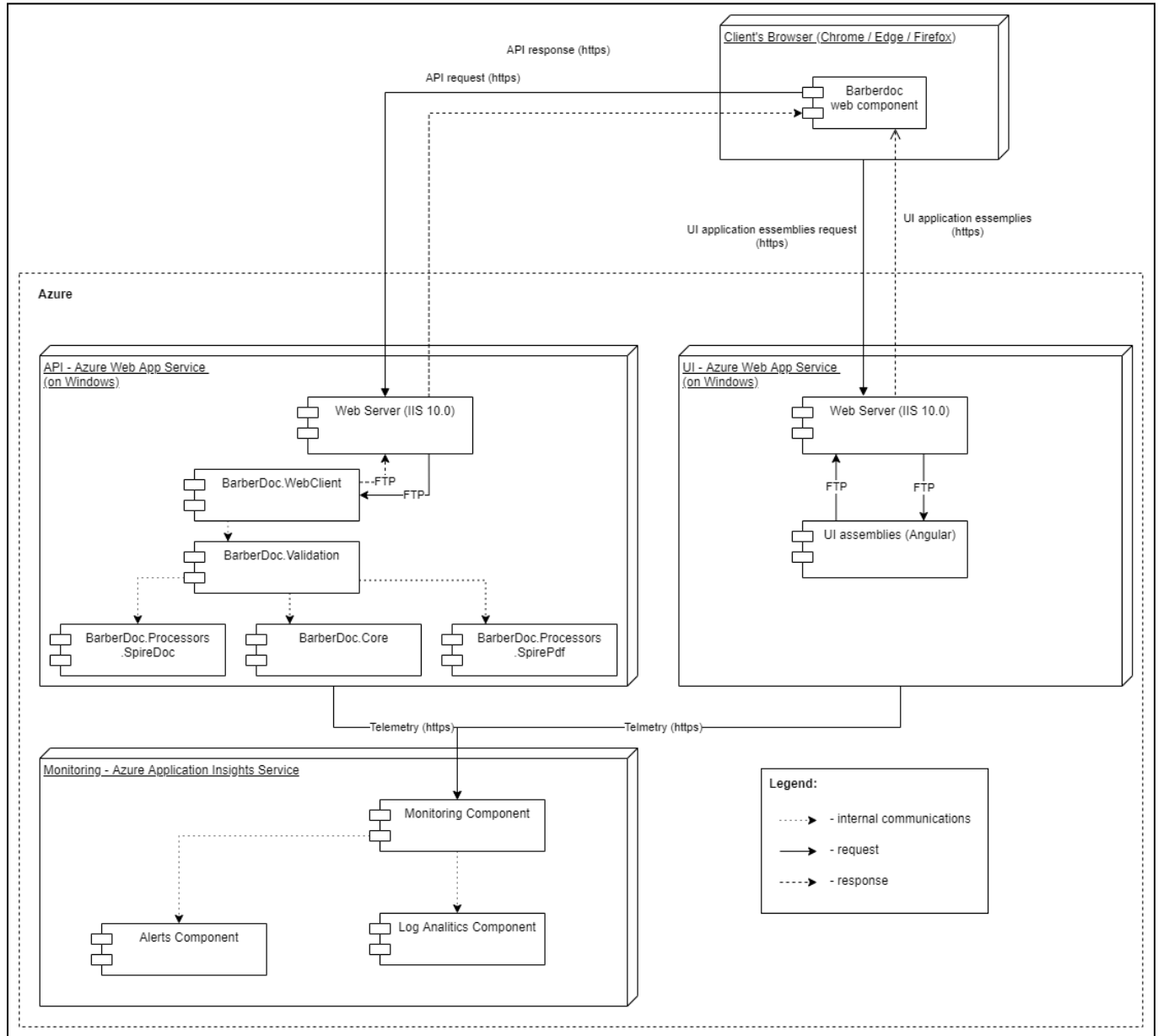
5 ANALYSIS MODELS

5.1 Sequence Diagrams



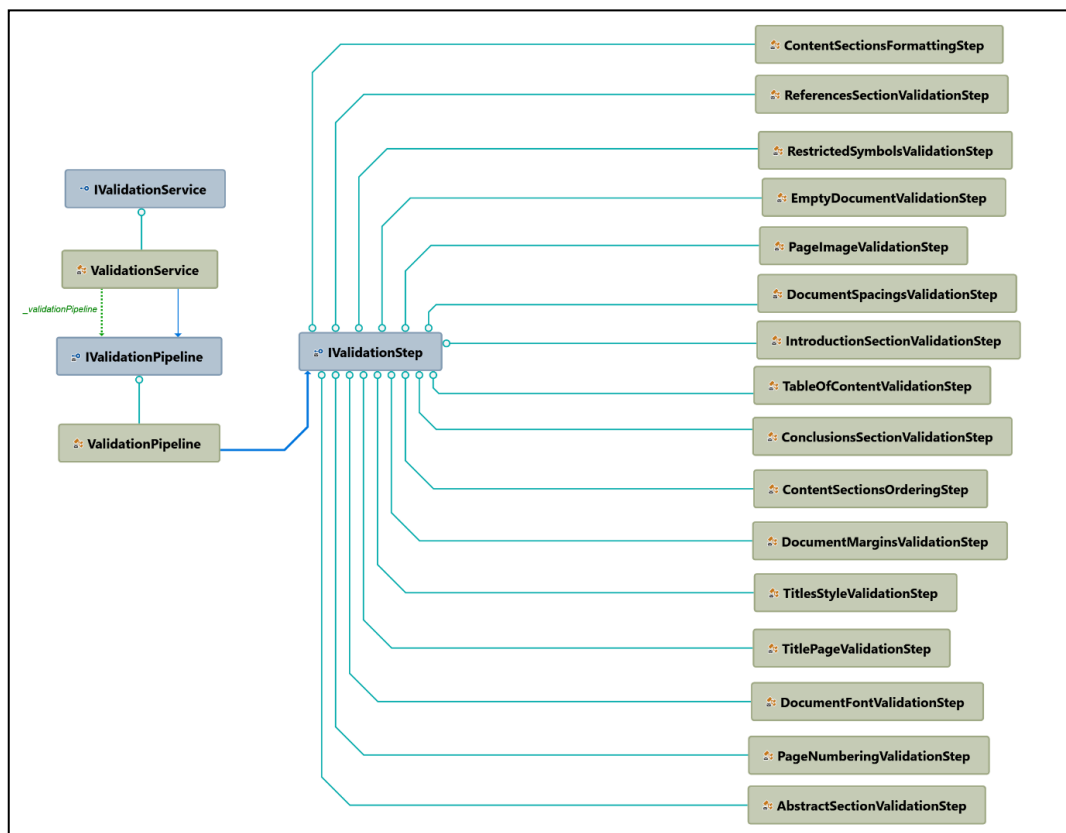
Picture 2 - Sequence Diagrams

5.2 Deployment Diagram



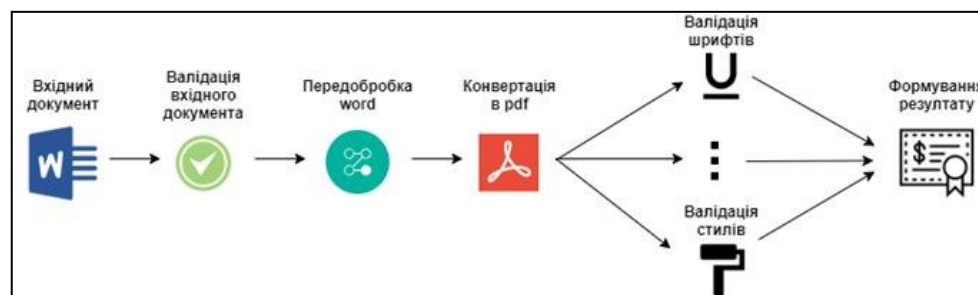
Picture 3 - Deployment Diagram

5.3 Class diagram



Picture 4 – Class Diagram

5.4 Concurrency diagram



Picture 5 – Concurrency Diagram

6 CHANGE MANAGEMENT PROCESS

Change management process consists of the following steps:

1) Request for change is made.

Changes and issues are captured through Project Team meetings, a working group of external stakeholders including other government agencies, meetings of the Project Board, outcomes of approval stages and through monitoring by the Project Team and Project Sponsor.

2) Register and assess the change.

Change requests are captured on a central change log that enables monitoring of change levels by the Change Manager.

3) Review and submit RFC to Change Board.

If the change is valid the Project Manager assesses the impact of the change with the project team and submits the change via a Request for Change Form to the Change Manager. It is then submitted to the Change Board who assess the change and approve or reject it.

4) Change Board accept or reject the change.

The Change Board includes the client and senior managers who have a strategic overview of the contract. They assess whether the change is within program tolerances and are able to approve most changes.

5) Update plans & implement the change.

Once the change is approved the change log is updated and stakeholders are informed. The Project Manager then works with the project team to plan the implementation of the change; amending the project and stage plans, sourcing any additional equipment or personnel, updating the configuration items and completing the work package(s).