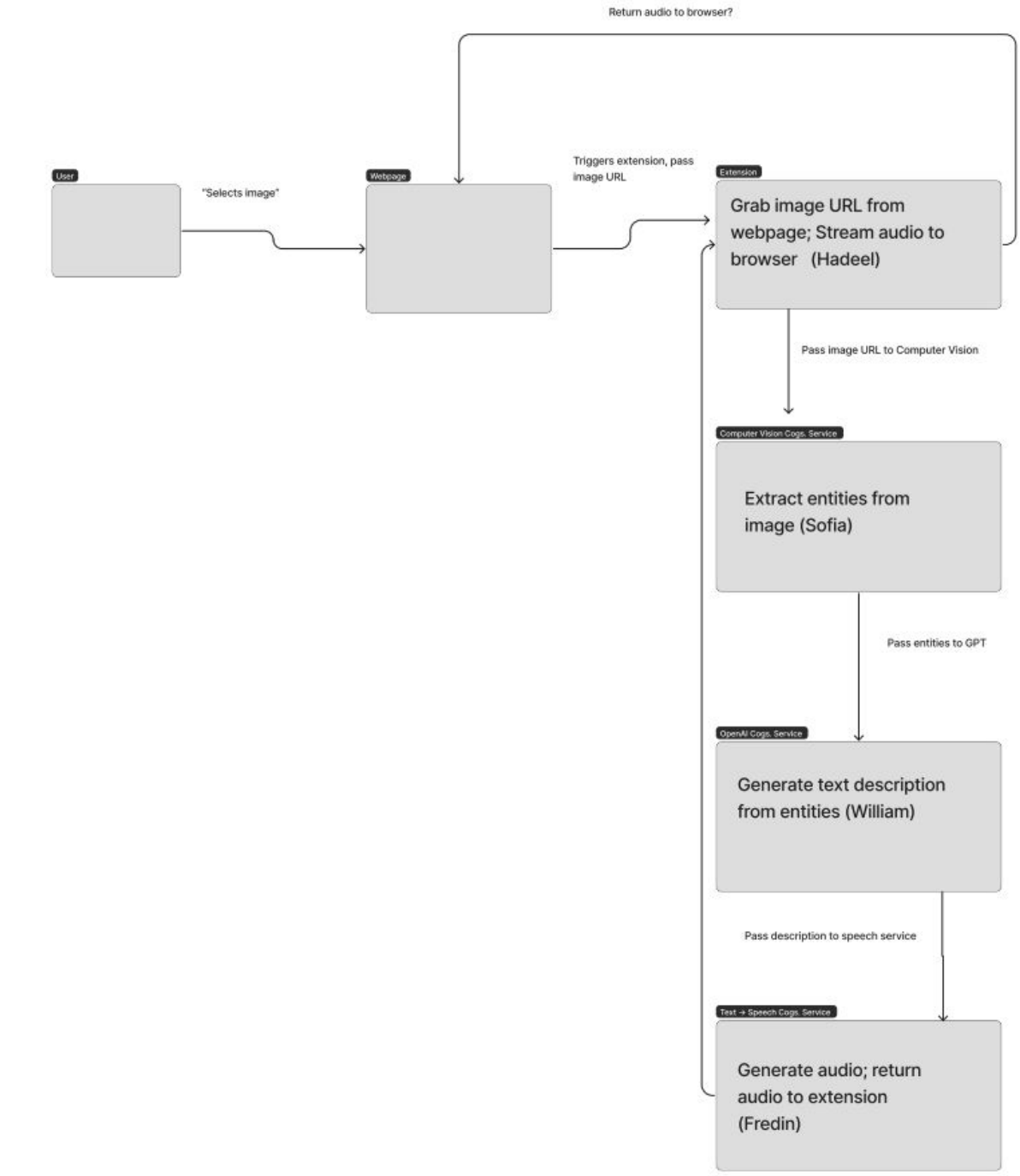
Dev Spec (Team OrcaBox)

Problem Statement

How can we increase the accessibility of visual media in browsers for people with visual impairments to create equal access to information and experiences on the web?

Architecture and Detailed Design

The solution design is for generating a description of an image using the Azure Computer Vision service and converting that description into speech using Microsoft Cognitive Service. The process consists of sending a JSON object with the URL of the image to the Azure Computer Vision service, which returns a JSON object with various properties, including captions, objects, labels and landmarks. This JSON object is then sent to a description generation service that uses an LLM (daVinci model) service to generate a description. The resulting description is converted to speech using the Microsoft Cognitive Service: text to speech. The audio file is uploaded to the Azure Storage blob service and a unique URL is generated for each uploaded audio file. Finally, the URL is fetched and returned as an endpoint response, and the audio is played.



A. Extension Integration (Hadeel)

- Utilizing the listeners event to get the URL images that the user select
- Play the audio after to receive it.

B. Azure Computer Vision Cog. Service (Sofia)

- Utilizing an Azure Function with .NET runtime
- Azure Function connects to Computer Vision Cognitive Services (V

C. Azure OpenAI Cog. Service (William)

- Utilizing a prompt to generate the template that generates the description with davinci Model
- Azure Function connects to OpenAi Chat GPT davinci

D. Azure Text To Speech Cog. Service (Fredin)

- Utilizing an azure function to connect to Azure Storage Blob to upload each audio.
- Azure Function connects to Microsoft Cognitive Service text to Speech

Sequencing, Integration, and Dependencies

Dependencies:

1. Azure Computer Vision Microsoft Cognitive Service
2. Azure OpenAi Microsoft Cognitive Service
3. Azure text to Speech Microsoft Cognitive Service

Sequencing:

1. The URL will allow us to extract the information from the image to create the description. Therefore, we will have to perform a fetch to the endpoint of the Azure Computer Vision service. To perform the fetch, we will need the URL of the image as a JSON object. Finally, the endpoint will return a JSON object with the following properties
    a. captions
    b. objects
    c. tags
    d. ocrText
    e. landmarks
    f. imageType
    g. ColorSchema
2. The JSON object will be sent to fetch the endpoint of the description generation service. In this case we will be using an LLM service (daVinci model) to generate a description, so we will implement a template that works as a prompt and generates the description. The endpoint will return a JSON object containing only one property, which is the generated description.
3. Finally, the last endpoint to fetch is the Microsoft Cognitive Service: text to speech.
    a. The conversion of the text to speech, with the configuration of an English speaker will be generated.
    b. In this way an audio is generated which cannot be stored locally because the reproduction of the audio for the user will be done using a URL.
    c. The next thing is to use the Azure Storage blob service to load the audios in that service.
    d. When the generated audios are uploaded a unique URL can be generated for each uploaded audio.
    e. Each URL is fetched and returned as the endpoint response.
4. The URL is saved in a variable and then it is used to generate an Audio object.

5. The audio is played.

Validation & Testing

Testing:

1. Test the functions that fetch eachpoint
2. Test the function that generate the image properties with Azure computer Vision
3. Test the template to generate different description using the model Chat GPT Davinci
4. Test the function that transform the text to Speech using text to speech of Microsoft cognitive service