Data flow problem in EO

Introduction

To effectively integrate AI into EO and enable individual chapters to fully leverage the potential of this technology, the issue of data flow within our organization must first be addressed.

Each chapter has two main types of data:

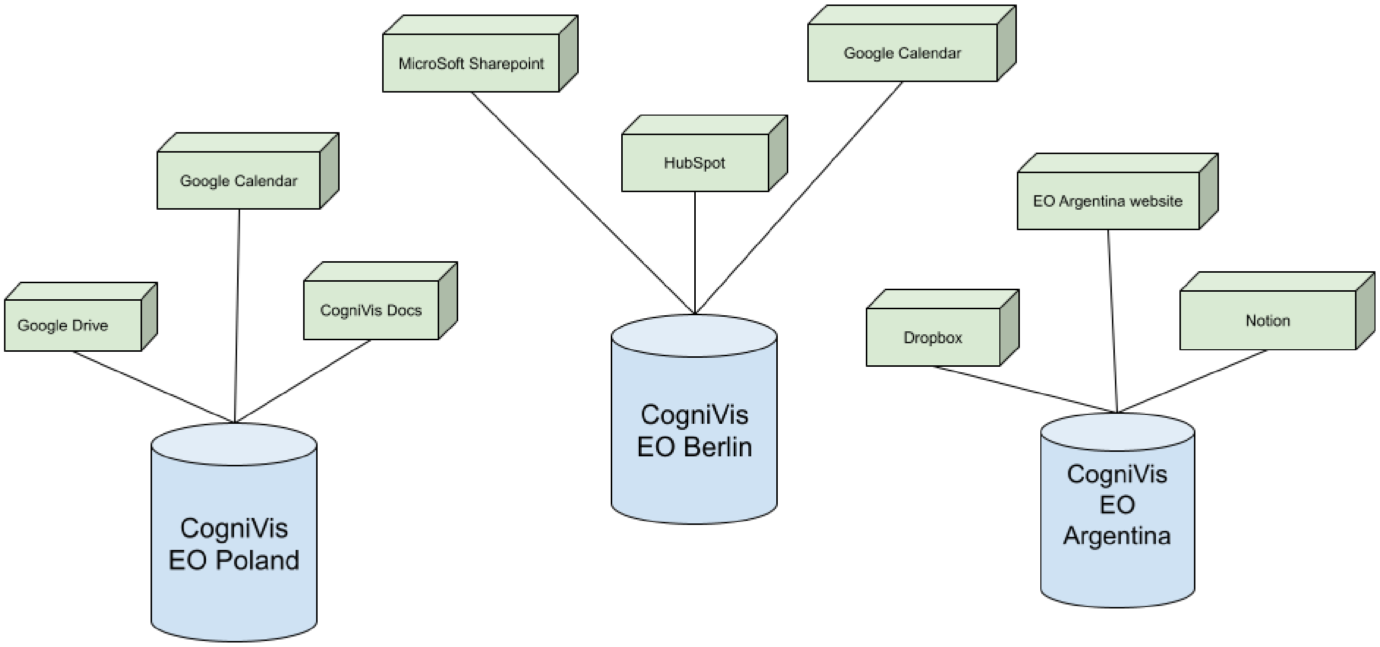
1. **Data specific to their chapter**, unique to that chapter and accessible only to its members – for example, information about chapter members, chapter budgets, internal chapter procedures, etc.
2. **Data shared across the entire EO** – such as official training materials for forums, guides for individual board members, ,branding materials, etc.

Both points require careful consideration and design, but in this discussion, I would like to **focus on point number 2** – data shared across the entire EO.

Current Data Flow Model

In the current AI solution for EO, we use the **CogniVis AI software**.

In CogniVis AI, we create a separate instance/unit of the software for each chapter. This ensures that each chapter has full control over its data and can freely manage user accounts for its chapter.



Link to the diagram above: https://docs.google.com/drawings/d/1PgJEkZtRAytCi81ziFLFp5_FqXDedBI6m-LUXUZshew/edit?usp=sharing

Legend:

1. **Blue cylinders** represent CogniVis instances for individual chapters (e.g., EO Poland, EO Berlin, EO Argentina).
2. **Green rectangles** represent connected data sources (so-called connectors) for each chapter’s instance. Each chapter can use different data sources – for example, EO Poland may use its Google Drive, EO Berlin its Microsoft SharePoint, and EO Argentina its Dropbox.

Adding Data to a Chapter Instance

Let's consider a **simple example**: **each chapter wants to add two files** to its instance so that AI can later use them to respond to questions related to these files:

1. The first file is a spreadsheet with data on the members of that chapter

Example spreadsheet with member data: https://docs.google.com/spreadsheets/d/1BbusZF1i6689Je_JOENt4arsVNTTC9phJ0NmznI51Ug/edit?usp=sharing

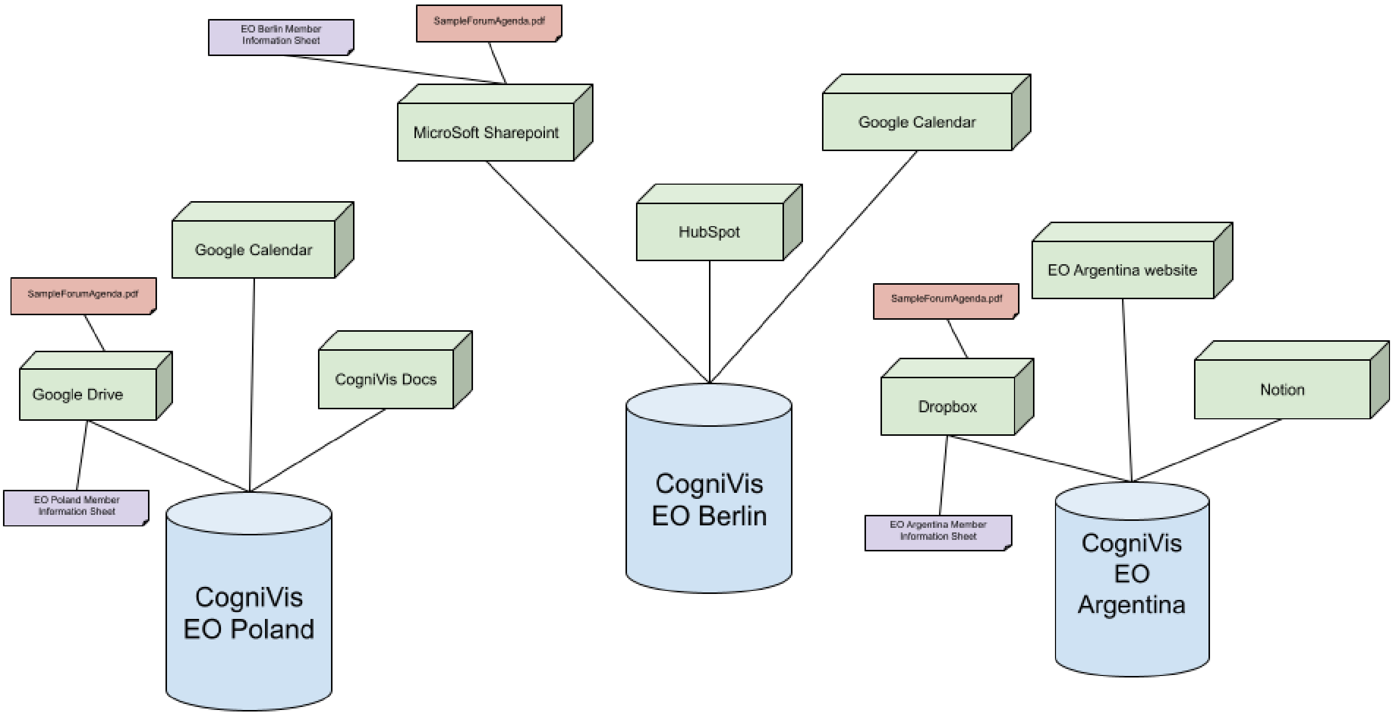
Each chapter will have such a spreadsheet separately, and each chapter wants its sheet to be accessible only to its members.

Thus, each chapter will add this spreadsheet (marked **purple** in the diagram below) to its data source. Following our example schema (see diagram below):

1. EO Poland will add the sheet "EO Poland Member Information Sheet" to its Google Drive.
2. EO Berlin will add the sheet "EO Berlin Member Information Sheet" to its Microsoft SharePoint.
3. EO Argentina will add the sheet "EO Argentina Member Information Sheet" to its Dropbox.

2. The second file is the PDF "SampleForumAgenda.pdf" which is an official document downloaded from https://www.eonetwork.org/

Again, each chapter will add this file "**"SampleForumAgenda.pdf**" (marked **red** in the diagram below) to its data source.



Link to the diagram above: https://docs.google.com/drawings/d/1bv84hB65vT7kwyh99-chpx1RZiR2TwXHu-xFXwqT3pM/edit?usp=sharing

Problem Analysis

In the above data flow schema, **it is correct that each chapter adds its member data sheet to its CogniVis instance**, as each chapter will have a different file, and access should be restricted within that instance.

However, it is **not optimal** that the "**"SampleForumAgenda.pdf" file is also added individually to each instance**, despite being identical and containing data shared by all EO chapters.

For example, if EO Global releases a new version of this file, all chapters will have to update it individually in their instances, adding a lot of maintenance work and creating risks, such as a chapter forgetting to update and using outdated versions of the official EO documents.

Moreover, the issue becomes more complex considering the large volume of official EO data and documents, and the continuous release of new ones. If each chapter has to individually update these files, data discrepancies will quickly emerge, leading to inconsistencies and, eventually, complete disarray, significantly reducing the effectiveness of the AI that relies on this data.

Solution and Suggested Data Flow Model

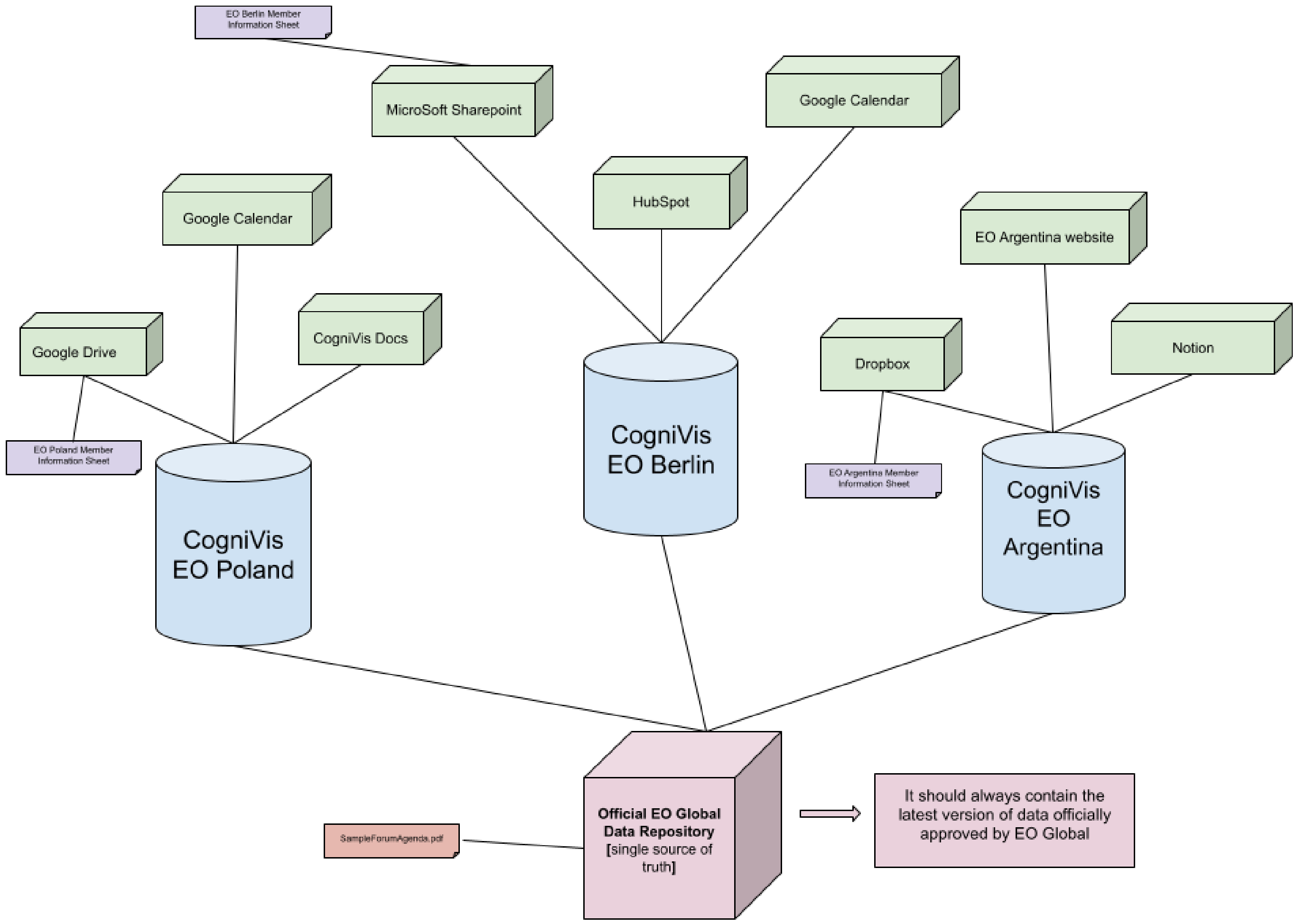
The data flow should be changed so that the **official documents and data from EO Global**, which are shared by all EO chapters, **have a single source** from which all CogniVis instances for all chapters can pull data.

In this scenario, **individual data** (such as member data sheets – marked in purple) **would still be added individually by each chapter to its own instance**.

However, **general data for the entire EO** (such as the "SampleForumAgenda.pdf" file – marked in red) **should be stored in a single, official EO Global data repository** that always contains the most up-to-date data.

Then, all CogniVis instances for all EO chapters could pull official global data from the EO Global repository, while adding their private data individually to their own instances.

This would significantly reduce the maintenance burden of shared data for the entire EO, as it would only need to be maintained and updated in one place.


EO Berlin Member Information Sheet
MicroSoft Sharepoint
Google Calendar
HubSpot
Google Calendar
EO Argentina website
Google Drive
CogniVis Docs
Dropbox
Notion
EO Poland Member Information Sheet
CogniVis EO Berlin
EO Argentina Member Information Sheet
CogniVis EO Argentina
CogniVis EO Poland
SampleForumAgenda.pdf
Official EO Global Data Repository [single source of truth]
It should always contain the latest version of data officially approved by EO Global

Link to the diagram above: https://docs.google.com/drawings/d/1t2FvtLyfs-qvEVp2gA4pqaPdub46hZOI9KgwzRY8gPI/edit?usp=sharing

What Exactly Should the EO Global Official Data Repository Be?

Below are some suggestions and considerations for possible solutions:

1. Cloud Storage

In the simplest solution, EO Global could set up cloud storage (Google Drive, Microsoft SharePoint, Dropbox, etc.) that would be regularly maintained and approved by the EO Global team.

2. API Communication with https://hub.eonetwork.org/

A more advanced solution would be enabling direct API communication between the CogniVis instances and https://hub.eonetwork.org/.

The question is **whether** the data on https://hub.eonetwork.org/ is regularly maintained and always contains the latest versions of all documents.

Issue with PDFs for Official EO Global Documents

CogniVis AI performs well in reading PDF files and, in most cases, provides correct responses based on them.

However, PDF files (the format of most EO Global documents) are not the ideal solution and create many long-term complications, such as:

- **Difficulties in data extraction**: The structure of PDF files is designed primarily for visual presentation rather than storing and processing information by machines. AI often encounters issues in correctly recognizing text, tables, graphics, and document layout, leading to errors in data extraction.
- **Lack of consistent structure**: PDF files do not have a unified standard for data layout. Even in similar documents, formatting may vary, complicating AI's interpretation of information such as headers, lists, or text sections.
- **Limited access to metadata**: Unlike other formats like JSON, XML, or CSV, PDF files do not contain structured metadata that can be easily analyzed by algorithms. This greatly limits the ability to search and filter information.
- **Character encoding issues**: PDF can store text in different encoding formats, which often causes problems in recognizing certain characters, especially in multilingual documents or when using non-standard fonts.
- **Inefficient processing of multi-page data**: AI algorithms may struggle to recognize context when content is spread across multiple pages. For example, sentences may break at the end of one page and continue on the next, leading to incorrect interpretations.
- **Lack of efficient and quick updates**: PDFs are generally static, making them unsuitable for dynamic updates and automatic data retrieval. For AI, this means manual updates to sources are required each time.
- **Challenges in recognizing images**: PDF files often contain text stored as images, which requires additional processing using OCR (Optical Character Recognition), not only lengthening the analysis process but also potentially generating errors, especially with low-quality scans.
- **Complicated semantic analysis**: AI has difficulty understanding context in PDF files since the text is often arranged in a nonlinear manner (e.g., in columns or inserted in

frames). This can lead to misinterpretation of context, meaning, and relationships between text fragments.

This is a problem to solve in the future (currently, even using PDFs, we can deliver a lot of value with AI for EO). However, the ultimate goal would be to devise a different solution. A document management system would be needed that would allow for the creation of an optimal structure for AI and easy updates.

Summary

We primarily need to find a solution for the EO Global Official Data Repository. Treat this document as the beginning of a brainstorming session and share your ideas for addressing this challenge in the comments.

Revision #9
Created 9 October 2024 07:12:51 by Admin
Updated 9 October 2024 09:25:57 by Admin