

Supporting humanitarian efforts in Afghanistan: Developing the ability to live-link different data sources to Power BI

Ojaswee Bajracharya



4th Year Project Report
Computer Science
School of Informatics
University of Edinburgh

2024

Abstract

This project is part of a larger initiative that aims to create a visual dashboard system for combining multiple humanitarian data sources covering various factors. The platform selected for this is Power BI, which facilitates data visualization from different sources. This particular project seeks to improve the data picture for crisis-affected countries such as Afghanistan by streamlining the process of linking data from large global sources into Power BI. A prototype was created to test the viability of establishing a connection with various data sources and to identify any difficulties associated with it. A user study was also conducted to test the functionality of the live link and identify other features that need to be included.

Research Ethics Approval

This project obtained approval from the Informatics Research Ethics committee.

Ethics application number: 773411

Date when approval was obtained: 2024-02-18

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Ojaswee Bajracharya)

Acknowledgements

I would like to thank my supervisors Fiona McNeill and Amanda Meyer for all their support and feedback throughout the project. I would also like to thank my friends for making the past four years truly memorable and my family for always being incredibly supportive.

Table of Contents

1	Introduction	1
1.1	Motivation	1
1.2	Hypothesis and Objectives	2
1.2.1	Hypothesis	2
1.2.2	Objectives	2
1.3	Chapter Structure	3
2	Background	4
2.1	Situational Background	4
2.2	Technical Background	5
2.2.1	Power BI	5
2.2.2	Extract Transform Load (ETL) Process	6
2.2.3	Extracting data from websites	6
2.3	Existing platforms in the Humanitarian Landscape	8
2.4	Data Sources	9
2.5	Chapter Summary	9
3	Design	10
3.1	Prototype Design	10
3.1.1	System Overview	10
3.1.2	“Live” in Live-Linking	10
3.1.3	System Design	10
3.1.4	Requirements	11
3.2	Global Data Sources	12
3.3	Data Extraction Feasibility	12
3.3.1	API Usability	12
3.3.2	Web Scraping	17
3.3.3	Downloadable Resources	18
3.3.4	Global Data Extraction Feasibility Overview	18
3.4	Prototype Data Sources	19
3.4.1	Indicators for prototype	19
3.4.2	Python Data Format	19
3.5	Power BI Design Implications	20
3.6	Chapter Summary	21
4	Implementation	22

4.1	Prototype Implementation	22
4.1.1	General Process	22
4.1.2	API data to Google Sheets Processes	23
4.2	Software Testing	26
4.3	Requirements Testing	27
4.4	Chapter Summary	27
5	Evaluation	28
5.1	User Study Design	28
5.2	End Users	28
5.2.1	General End Users	28
5.2.2	User Study End User	28
5.2.3	NASA-TLX Evaluation	29
5.2.4	User Study Process	29
5.2.5	Materials	30
5.3	User Study Results	30
5.3.1	User Study Set Up	30
5.3.2	Mid-study Semi-structured Interview Findings	31
5.3.3	Mid-study Changes Findings	31
5.3.4	Post-mid-study modifications	32
5.3.5	Post-Study Semi-Structured Interview Findings	33
5.3.6	Post-Study Changes Findings	34
5.3.7	NASA-TLX Findings	34
5.3.8	Overview of User Study	34
5.4	Chapter Summary	35
6	Conclusion	36
6.1	Discussions	36
6.1.1	Live Linking Discussion	36
6.1.2	Power BI Discussion	37
6.1.3	Lessons Learnt	37
6.2	Limitations	38
6.3	Future work	39
6.4	Conclusion	40
	Bibliography	41
A	Participants' information sheet	46
B	Participants' consent form	49
C	NASA TLX	51
C.1	NASA TLX Scale	51
C.2	NASA TLX Workload Comparison Cards	51
D	Semi-structured Interview Questions	55
D.1	Mid-study Semi-structured Interview Questions	55

D.2	Post-study Semi-structured Interview Questions	55
-----	--	----

Chapter 1

Introduction

1.1 Motivation

Data collection in crisis-affected countries is rarely the highest priority, yet it is essential for improving situations in these countries [Montjourides, 2013]. Data plays a pivotal role in planning, coordination, and monitoring efforts and has the potential to make a difference in policy-making. Additionally, on a global scale, the systematic collection of such data serves many purposes such as preventing human rights violations against children [Devries et al., 2016], supporting the tracking of development goals [Group et al., 2006], and heightening the international awareness of various pressing issues [Madsen et al., 2016]. The consequences for children in these countries due to the lack of education include their prospects for employment and personal growth dwindling significantly. This can lead to significant consequences for the development of the country. The results of the project could result in more accurate assessments of the situation and enable appropriate funding allocation by ensuring decision-makers have access to the most current data from diverse sources, allowing them to get a comprehensive picture. This project focuses in particular on education in Afghanistan. Research has shown that education during conflict is also a priority of local communities [Buckner et al., 2022]; therefore, the results of this project will help allocate more resources to the areas that need it.

In 2021, there was a change in the Afghanistan government as the Taliban assumed control, adding challenges in data collection, and hampering the ability of humanitarian organizations to address urgent needs. Before this transition of power, Afghanistan had made considerable improvements in increasing the enrolment of girls in both primary and higher education, while also achieving significant progress in literacy rates among boys and girls, however, there was still a lot of progress to be made [UNESCO, 2021]. Following the change in leadership, new legislation such as prohibiting girls from attending high school has hindered this progress for girls [Save the Children]. Even before the Taliban's takeover, the data was not as advanced as it could be, with the focus being on a small number of indicators, with the national data system, the Education Management Information System (EMIS), being in the early stages of development [Saraogi et al., 2017]. Often, global organizations employ distinct data-

collection systems and gather information at varying points in time, giving rise to challenges related to data scarcity and its lack of disaggregation [UNESCO Institute for Statistics, 2017]. Currently, charitable organisations encounter challenges when navigating humanitarian data systems as humanitarian data is often spread out across multiple systems and platforms. Individuals have highlighted the necessity for a tool with the capacity to facilitate data creation, sharing, utilization, and analysis [Buckner et al., 2022].

This project exists as part of a wider project (Education Data in Fragile and Conflict-Affected Countries Project) which I will refer to as the parent project. The objective of this parent project is to develop a visual dashboard system on Power BI which combines data from global, national and ad-hoc datasets along with primary research on a wide range of factors including education outcome indicators, health, protection, climate etc. This system aims to provide a solution to improve the data picture for countries in crisis by providing a single platform for all relevant data, maximising data sharing and collaboration. The final system will be built on Power BI and therefore, this specific project will focus on streamlining the process of linking data from large global sources into Power BI and exploring the feasibility of establishing real-time connections with multiple data sources. Ultimately, this effort aligns with the broader project's mission of advancing Sustainable Development Goal 4, which is to ensure access to quality education for all children [United Nations, 2023]. Creating a live link will help ensure the data is up to date hence allowing for a more accurate assessment of the situation.

1.2 Hypothesis and Objectives

1.2.1 Hypothesis

This project aims to assess the viability of developing a system that can establish live connections to a range of data sources within Power BI. This project seeks to significantly improve data timeliness, allowing users to have the most up-to-date data available. This leads me to the following hypothesis:

Hypothesis: Live linking different data sources to Power BI would allow users to more effectively access relevant data.

1.2.2 Objectives

To meet the aims of this project I will:

- Research live-linking systems or other similar systems that exist both in the humanitarian sector and other contexts
- Design a system to live-link a range of data sources to Power BI
- Implement a prototype of the live-link system and test the prototype through software testing

- Design an evaluation methodology tailored to assess the practicality of the system under conditions where updates are irregular
- Evaluate using the designed evaluation approach
- Suggest areas to look into for future work

1.3 Chapter Structure

Chapter 1 - This chapter delves into understanding the wider situational context and explains the parent project. It provides an introduction to the problem being solved in this project and the hypothesis.

Chapter 2 - This chapter presents the contextual background of the situation in Afghanistan and the importance of reliable and up-to-date data. This chapter also introduces a background to the systems and the methods used in solving a problem such as this.

Chapter 3 - This chapter describes the design stage for the creation of a prototype that links different data sources to Power BI.

Chapter 4 - This chapter describes the implementation process and the software testing carried out.

Chapter 5 - This chapter outlines the evaluation methodology and presents the prototype evaluation results.

Chapter 6 - The final chapter looks at the results of the project on the whole and whether it answers the research questions. This chapter also looks into some discussions, limitations and future work.

Chapter 2

Background

2.1 Situational Background

Fragile and conflict-affected countries such as Afghanistan face numerous challenges regarding data collection and management due to security and logistical constraints and travel constraints [DFID - GOV, 2010]. Global Non-Government Organisations (NGOs) and global organisations such as UN agencies collect data to allow for effective development interventions to be put in place [Hoogeveen and Pape, 2020]. Regarding global NGOs, data is often collected on the ground through face-to-face household interviews and this has led to a mixture of independent data collection efforts by a range of different organisations [REACH, 2022]. Consequently, various organisations have different data for the same regions which complicates the mission of having a unified and accurate understanding of the situation. This conflicting data is worsened by the lack of a robust centralised data system which accentuates the challenge of collating, analysing and interpreting data.

The lack of reliable data causes significant issues in the aspect of planning within the development landscape. Planning is often carried out under pressing, ad-hoc conditions, emphasising the need for up-to-date and trustworthy data. Data also plays a pivotal role in determining the allocation of government funds and identifying the areas of critical humanitarian need. It also serves as a resource for monitoring and evaluating the evolution of the situation [Buckner et al., 2022]. In other countries, like South Sudan, the Ministry of Education, Science and Technology worked with UNICEF to conduct a rapid assessment of the country's education system. This prepared them for the development of an Education Management Information System (EMIS) and since then, South Sudan has made significant progress in the years after its independence. Without the use of such tools, assessing progress would be very difficult [Montjourides, 2013]. Afghanistan has an emerging EMIS with a focus on a small number of indicators [Saraogi et al., 2017] and data collection from agencies normally takes place when they evaluate their projects and satisfy the needs of their donors [Shalash et al., 2022]. Therefore, the establishment of an up-to-date centralised system covering a large number of factors and collating the most recent information from various organisations could be extremely beneficial for the development of education in Afghanistan [Saraogi et al.,

2017].

Different organisations are also focused on separate statistics. For example, UNESCO is focused on certain education statistics: education, science, culture and communication and does not include statistics such as child marriage. Furthermore, UNESCO in particular is also involved in generating statistical estimates for statistics with very little data such as out-of-school data [UNE, Accessed 2024b]. However, these organisations do not include data from smaller surveys and ad-hoc research and are more focused on collecting data from multiple countries. Data from these organisations are often published on their websites but use different layouts. This makes it difficult for users to find information and standardise their analysis processes. For instance, UNICEF provides its data through the UNICEF data warehouse [UNICEF, Accessed 2024b] – a web application allowing users to filter data by a range of indicators such as education, nutrition and more. Within these categories, there are sub-sections which represent the indicators e.g. “Completion rate for children of primary school age”. Navigating this sub-section leads to a comprehensive table with data for all countries regarding the chosen indicator. Users can then refine their search to filter to only Afghanistan, which has data for that particular indicator that you can view and download [UNICEF, Accessed 2024b]. Similarly, UNESCO displays its data through the UNESCO Institute for Statistics, which offers a data browser for viewing and downloading data in a table format. [UNE, Accessed 2024a]. The World Bank also offers a Data Bank similar to UNESCO’s and UNICEF’s applications and also an Open Data catalog to access their data [World Bank Data Bank, Accessed 2024] [World Bank, Accessed 2024b].

It is also worth noting that data also has large gaps. For example, the indicator “Adjusted net attendance rate for children of primary school age” was updated in 2015 in UNICEF. The data is updated very infrequently and it is also unpredictable when the data will be updated. The consequences of a lack of timely data could lead to the inability to address emerging issues promptly and adds challenges to monitoring and evaluating the situation. Thus, for someone looking for information on Afghanistan’s situation, the task becomes complicated. It requires sifting through various separate data sources without knowing when there will be a new update.

2.2 Technical Background

2.2.1 Power BI

The overarching system will be built on Power BI, a platform which consists of a collection of services, apps and connectors used to bring together different sources of data to create analytics and visualisations [Power BI, Accessed 2024d]. Power BI in general allows users to connect diverse data sources ranging from databases, cloud services and web services. This platform also allows for data cleaning and transformation, data modelling and the creation of interactive reports and dashboards. Moreover, developers can also build extensions allowing for more advanced functionality. There are two main components of Power BI important to this project: Power BI Desktop and Power BI Service. Power BI Desktop is a free application that can be downloaded locally. Power BI Service is a cloud-based platform which extends the analytics and

visualisation capabilities to the web, allowing people to be able to collaborate and work on the dashboard in real-time [Power BI, Accessed 2024b]. Power BI has been chosen as the platform to use because it is user-friendly and can be used both by non-technical and technical users.

Multiple other humanitarian organisations have also used Power BI. For example, the Education Cluster and UNICEF have developed a blueprint for Management Information Systems to do with Community-Based Education [Afghanistan Education Cluster Dashboard, Accessed 2024]. They have a dashboard for Afghanistan which is made with Power BI however it is static and only provides the most recent data and no ability to see previous months or years. The UN Displacement Tracking Matrix also has a global survey to prevent their operations and activities and this is presented using a Power BI dashboard [UN Displacement Tracking Matrix, Accessed 2024]. Again, it only provides data for that given year and does not include past data from previous years.

2.2.2 Extract Transform Load (ETL) Process

Extract Transform Load (ETL) is a process used to load data from multiple sources into a data repository [IBM, Accessed 2024a]. This involves extracting data from different sources, cleaning the data into a form that is usable and loading it into their desired system. If this were in a business context, a process such as this could have been used to connect the data to a database. However, creating such data pipelines can be challenging as they are built with complex custom code which results in limited reusability. There are tools to help with the steps in the processes and also the option to custom code it.

2.2.2.1 Processes

- **Extract** - The extraction process involves pulling data from a range of data sources such as APIs, sensor data, marketing tools etc. These systems often also return data in varying data types ranging from structured JSON data to more custom-based outputs.
- **Transform** - The transform process uses a schema to alter the data. This can involve cleaning the data and validating and authenticating data. It can also involve steps such as formatting the data, performing calculations or removing data.
- **Load** - The load process is usually an automated process which involves moving the data into a data storage area.

2.2.3 Extracting data from websites

The two most common ways data is primarily extracted from websites is through the use of Application Programming Interfaces (APIs), or web scraping. These techniques serve as a foundation of data gathering, allowing us to access a range of information from online sources.

2.2.3.1 Application Programming Interfaces

An Application Programming Interface (API) is a way to allow different software applications to communicate with each other [Amazon Web Services, Accessed 2024]. They can be used for many purposes; for example retrieving data from a server, accessing specific resources or performing particular actions in a software. APIs define the methods and formats developers can use in their applications to request information. They are usually also documented to provide information to developers on how to use them with information about their endpoints, formats, response structures and more.

In the Business Intelligence industry, analysts often need to aggregate data from heterogeneous sources [Awasthi, 2012]. Tools similar to Power BI such as Tableau and Qlik provide features to allow users to connect to a diverse range of APIs and other data sources [Tableau, Accessed 2024] [Qlik, Accessed 2024].

Overall, the main challenge to linking multiple different APIs is the issue of API's having widely different structures and response formats. For each API, a developer would need to look into the documentation to learn how to use the API and retrieve the relevant data from the response.

2.2.3.2 Web Scraping

Web scraping is an automated process of extracting data from websites. It involves fetching web pages, analysing the HTML content and then extracting the relevant data from it. While it is a useful means of data collection, it must be carried out with adherence to legal standards. [Zhao, 2017]

There are numerous paid web scraping tools available, primarily used by businesses for market research and data analytics. Nonetheless, there is a range of libraries that allow developers to create their own web scrapers. Python, in particular, serves as an ideal language for this purpose because it has libraries such as BeautifulSoup and Scrapy. [Sirisuriya et al., 2015]

However, web scraping is also subject to legal regulations that differ depending on the country and the website. Typically, the “terms of service” or “terms of use” provide explicit guidelines for website usage, and they may prohibit web scraping activities. Websites might also use a “robots.txt” file to indicate which sections can be scraped and which are off-limits. These legal measures are primarily aimed at protecting the website's intellectual property and addressing privacy concerns. [Zhao, 2017] Furthermore, excessive web scraping can strain a website's resources, leading to high levels of traffic that could slow down a website or crash its servers. Consequently, when taking on web scraping projects, it's important to be mindful and restrict the usage of web scrapers. Live linking requires sending requests regularly, which can strain the target website's resources.

One of the key challenges regarding web scraping is every website will be laid out differently so trying to web scrape a range of websites is difficult. The data will be stored in different places in different formats. Websites can also undergo updates, changing the websites' structure and affecting the scraping process and being able to

maintain scrapers to adapt to change is another challenge. Some websites also load data dynamically via JavaScript [Nixon, 2012], requiring scrapers to interact with the page and wait for content to load. This adds more complexity to the scraping process.

Efforts have been made to establish a framework for cross-website scraping. DI-ADDEM, for instance, represents an automated system for data extraction across various domains. The systems combine automated website exploration, and the identification of relevant data and generate data wrappers. DIADDEM overcomes the challenges of scraping by combining phenomenological and ontological knowledge. [Furche et al., 2014] However, this approach may not be directly applicable to our context, as the data we seek is not always present on websites e.g., it may exist as downloadable CSV files.

2.2.3.3 Areas utilizing regular website data extraction

Other groups of people facing similar challenges include bioinformaticians. They make extensive use of APIs to find and extract biomedical resources, however, for websites which do not offer such services, a scraping-like solution is required. There are frameworks and tools built specifically for bioinformaticians to use for web data scraping for their use cases. [Glez-Peña et al., 2014] A tool/framework for the particular use case the project will be assessing, has not yet been created.

2.3 Existing platforms in the Humanitarian Landscape

There are several platforms which bring together data from diverse sources, although these do not have a live-linking functionality. There are also limitations, particularly regarding the availability of Afghanistan data, which thus makes them less valuable for the context of this project.

2.3.0.1 Humanitarian Data Exchange

The Humanitarian Data Exchange (HDX) is a platform run by OCHA's Centre for Humanitarian Data, whose aim is to make humanitarian data easy to find and use by creating this platform to share data across organisations. It brings together open data, data which can be freely used by anyone and is open-source. Users can create accounts to share data (organisations require confirmation before posting) and the platform also makes use of scraper bots [UN-OCHA Humanitarian Data Exchange Project, Accessed 2024] which utilize APIs to gather data from organisations [Humanitarian Data Exchange (HDX), Accessed 2024a].

2.3.0.2 UNHCR - Operational Data Portal (ODP)

The Operational Data Portal, developed by UNHCR, allows information and data regarding refugee emergencies to be shared. They use "situation views" to cover different aspects of the refugee crises e.g. movements, returnees, camp coordination and management. The platform includes data from UNHCR reports and other sources, including government data. The website provides updates as recent as last month but

does not mention live-linking functionality [United Nations High Commissioner for Refugees (UNHCR), Accessed 2024].

2.3.0.3 Inter-Organizational Data Sharing and Collaboration

Global organisations often obtain data from other organisations. For example, UNESCO also makes use of UNICEF’s Multiple Indicator Cluster (Surveys (MICS) (UNICEF’s global household survey programme) [UNICEF Multiple Indicator Cluster Surveys (MICS), 2015]. Save The Children also draws data from international organisations such as UNICEF, WHO, UNESCO and World Bank [Save The Children, 2023]. However, each organisation prioritises different areas to focus on. For instance, UNESCO may not cover factors such as health and child marriage however these are factors that could be relevant to the context of the project.

2.4 Data Sources

For this particular project, there are a range of different data sources available, ranging from websites and PDFs to newsfeeds. The project will utilise a dataset pulled from the parent project. I will refer to this table as the “Data Sources Reference Table”. The data sources fall into three main categories: Global, National and Ad-Hoc. Global data sources refer to large organisations such as UNESCO, UNICEF, World Bank etc., national sources consist of data coming from government sources and ad-hoc data involves data from smaller organisations.

2.5 Chapter Summary

This chapter has looked into the challenges faced by humanitarian aid workers in terms of collecting data in crisis-affected countries, due to the information being scattered across various sources. This chapter also explores the features offered by Power BI, the platform the system will be built on, and different methods of extracting relevant data from websites. It also looked at what similar systems already exist in the humanitarian field.

Chapter 3

Design

3.1 Prototype Design

3.1.1 System Overview

The prototype will be a system primarily written in Python. Python has been chosen because some of the data sources provide Python libraries and it is suited for scripting shorter scripts. Since the platform needs to check for updates regularly, the system must be somewhere where this task can be performed daily. Ideally, this would involve running the system on a server that operates continuously, allowing for the set-up of a CRON job (a way to schedule and automate tasks on UNIX systems) [Hostinger, Accessed 2024]. However, due to the limitations of not having access to such a server and the cost associated with running it on the cloud, I have found a workaround for running the system from my laptop at a scheduled time for the prototype, using Windows Task Scheduler.

3.1.2 “Live” in Live-Linking

For this prototype, live has been determined to mean daily updates rather than real-time data synchronization due to feasibility and practical reasons. Real-time data offers access to the most current data however the continuous monitoring required for this would require resources such as servers to keep it running. Furthermore, for the intended purpose, data accuracy to this level is not essential and daily updates still offer the timeliness desired along with efficient use of resources. Therefore, providing data on a daily update basis still meets the project’s needs of delivering timely insights whilst being resource-sustainable.

3.1.3 System Design

The process of running the system will be a simplified ETL process. There is a prior preparation stage which involves getting the data onto Power BI Desktop and creating the schemas/tables etc. Once this stage is ready, the rest can be automated through the Python script and the refresh schedule on the Power BI Service.

The extract stage consists of getting the data from the data source. The transform stage consists of cleaning the data and formatting it to the form needed for Power BI and the load stage is loading the data onto the chosen database (in this case Google Sheets). Power BI's API is unable to automate the connection to Google Sheets therefore the process is split into two ETL processes. The first is the ETL process of data from the organisation into Google Sheets - extracting data from the organisation, transforming and cleaning the data and loading it onto Google Sheets. The second process is the ETL process of data from Google Sheets onto Power BI Service - extracting data from Google Sheets, transforming it to the correct data types in Power BI Desktop and loading it onto Power BI Service.

3.1.4 Requirements

The following requirements have been decided for the prototype system to have.

Functional Requirement	Justification
The system should connect multiple global web data sources to the Power BI system.	This is one of the main goals of the project and allows for a more comprehensive data picture.
The system should be timely and show the latest updates from each of the data sources (daily). The latest updates here include new data records and updates to existing records.	This satisfies the goal of the project being <i>live</i> linked and allows users to have access to the most current information, improving the accuracy of the data picture.
The system should make it clear what has been updated when a data source has been updated.	This will help users track changes in the data and allow them to take any actions required as a result of this.
The system should make clear when the data was last checked for updates.	This provides transparency and enables users to assess the accuracy of the data given when it was last updated.
The system should display the connectivity status to the different data sources.	This allows users to know when a particular data source is not currently connected and so the data might not be up-to-date.
The system should be accessible from anywhere with an internet connection.	This allows a range of users to be able to access easily and from anywhere and on any device.
Non-Functional Requirement	Justification
The system should be robust. It should be able to handle issues such as not being able to access APIs, APIs not returning correct responses, API's timing out.	Websites can go down and network connections can drop so the system must not break as a result of this.

3.2 Global Data Sources

There are 29 different global data sources listed in the Data Sources Reference Table (the main sources of data I will be focusing on for this project), with 28 of these sources being websites. Global data sources in this table refer to sources which are generally larger-scaled and collect data in several different countries. The main issues with these global data sources are the fact they rely on national data, infrequent household surveys or modelled estimates. This data can sometimes be out of date and incomplete and government data, for example, could be inflated which overall gives a false representation of the situation [Saraogi et al., 2017].

The feasibility of data extraction from these data sources also varies significantly. This variance usually comes from factors such as data availability due to accessibility issues to carry out more surveys and interviews [Kreutzer et al., 2019]. For this initial prototype, I will be focusing on getting data to be live-linked to power BI from web-based sources. Starting with a basic prototype, if this live link is possible, the prototype will provide a strong foundation to explore if other forms of data extraction to Power BI can be added to the system in the future.

3.3 Data Extraction Feasibility

In the global dataset, 13 data sources had APIs available with one in beta mode and one with a broken API (during the duration of this project). 5 offered an alternative form of downloading data (e.g. bulk download datasets, downloading as a CSV). The full list of what the different organisations offered can be found in Table 3.1. Due to the variation in the different forms of being able to access the data, I identified two main routes of obtaining this data in an automated way - APIs and web scraping.

3.3.1 API Usability

The usability of APIs among the available data sources varied significantly. Some organisations provided sophisticated APIs along with wrappers in the form of libraries or web interfaces to help generate API queries. For instance, HDX and Our World In Data (OWID) offered a Python API in the form of libraries [Humanitarian Data Exchange (HDX), Accessed 2024b] [Our World in Data, Accessed 2024]. In the instance where API's were not provided in the form of libraries, API calls were executed via HTTP. Some organisations such as UNICEF, OECD and UNStats had websites which allowed users to conveniently select multiple desired indicators to generate their API call. Most organizations that provide an API also offer documentation to assist users in utilizing its features. However, the ease of using APIs varied significantly among the different access methods available.

3.3.1.1 API Structure

For the APIs in these sources, they mainly use REST API. REST API communicate through HTTP requests and uses different types of HTTP request methods (POST, GET,

Table 3.1: Organisations and what they offer

Global Organisation	What it offers
UNESCO	Bulk Data Download Service
World Bank	API and Python API Library
UNICEF	SDMX API
OCHA (HDX)	Python API Library
Our World In Data	Python API Library
Save the Children	Nothing
EGER	Nothing
ACLED	API
CTDataCollaborative	Nothing
INFORM risk (EU)	API
OECD	SDMX API
SDG Dashboard	API
Fragile State Index	Downloadable CSV files
UNStats	API
UNPopulation	API (broken)
UN Human Development	API
UN Population Fund	Nothing
Uppsala Conflict Data Program	API
Vision of Humanity	Nothing
WHO	API
World Food Program	Nothing
ILOSTAT	SDMX API, Bulk Data Download Service
TRACE	Nothing
IDMC	Downloadable CSV files
UN migration DTM	Nothing (uses Power BI)
ND Gain Index	Nothing
IGME	Nothing
Food and Agriculture Organisation of the United Nations DIEM	Downloadable CSV files

PUT, DELETE) to carry out functions such as create, read, update and delete [IBM, Accessed 2024b]. API calls made through HTTP usually follow a specific structure involving an endpoint URL, an HTTP method, headers, an optional request body and optional query parameters [IBM, Accessed 2024b].

Since only data extraction is required in this project, GET requests with the endpoint URL and HTTP method are sufficient. The main sections of the API call used are the endpoint URL, headers and optional query parameters. An example of the structure for many of the API calls can be found in Figure 3.1. This type of call will return the data for the given indicators for that organisation in format specified.

`https://<endpoint_url>.com/<indicators>/<query_parameters>?format=json`

1 2 3 4

Figure 3.1: An example of the API call structure. 1 represents the specific URL endpoint to which the request is sent, 2 represents where the indicator code goes, 3 optional parameters used to filter data, 4 shows the data will respond in JSON form

3.3.1.2 Statistical Data and Metadata eXchange (SDMX)

Statistical Data and Metadata eXchange (SDMX) is an ISO standard to describe statistical data and metadata, designed to improve the sharing of data across similar organisations [sdm, Accessed 2024]. It provides statistical guidelines, technical standards and an IT architecture and tools. It is mainly used by governmental and international statistical organisations such as UNICEF, OECD and the International Labour Organisation.

The API call format is similar to the previous section. It starts with the base URL endpoint: “https://sdmx.data.unicef.org/ws/public/sdmxapi/rest/data/” and then adds the indicator codes separated by “+” before specifying the data format (e.g. json).

e.g. `https://sdmx.data.unicef.org/ws/public/sdmxapi/rest/data/UNICEF.AFGHANISTAN_CO,AFG_CO,1.0/.ECON_GVT_EDU_EXP_PTEXP+ED_ANAR_L1+HVA_PED_EID_NUM...?format=sdmx-json`

Furthermore, the websites which use SDMX come with a query generator to help create this API query. These generators allow you to select more than one indicator to extract data from, from one query. The structures between organisations where SDMX is used are also similar therefore, the same or a very similar script can be used for these organisations to extract data.

3.3.1.3 Issues with APIs: Inconvenient Set Up

When using APIs to obtain data, it is important to identify the specific indicators we want data for. In API calls, indicator codes are used to reference these indicators and these codes can vary for each website. However, where to find these indicator codes may not be immediately obvious. For example, the World Bank offers a DataBank and an Open Data tool to help find indicators [World Bank, Accessed 2024a]. For DataBank, inside the metadata information for the chosen indicator, the code can be found in the

title (see Figure 3.2). The indicator code can also be found in the ID of World Bank's Open Data feature which is a page that allows you to search for a specific indicator and provides graphs and comparisons across different countries. Thus, collecting indicator codes for multiple indicators to link can be time-consuming.

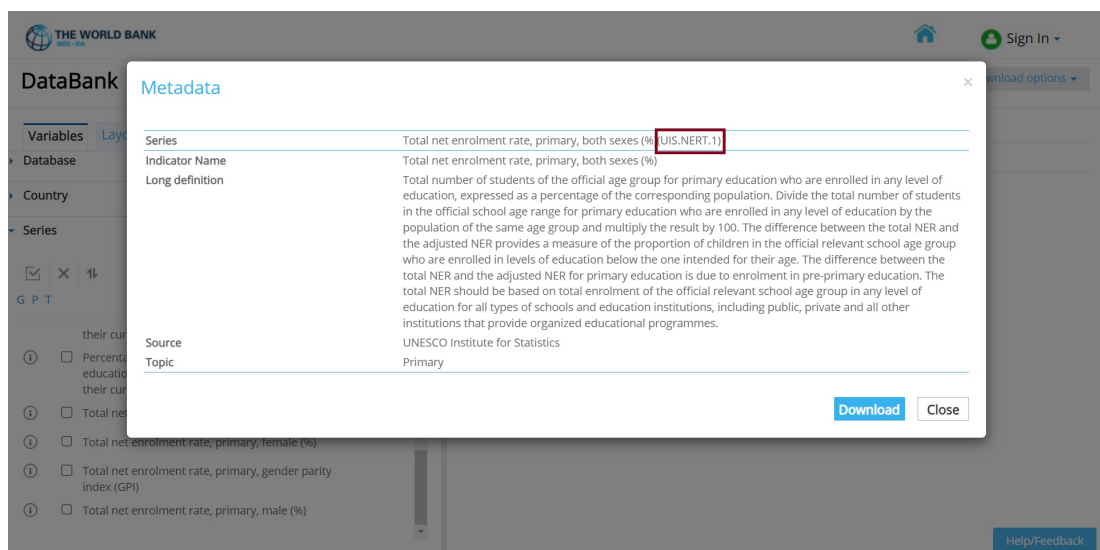


Figure 3.2: The indicator code in World Bank's Data Bank

Some organisations provided a website to help generate or test the query for the API call. For example, UNICEF's Indicator Data Warehouse, OECD Data Explorer and DRMKC's INFORM Web API Tester. An advantage of this website is you can request data from multiple indicators in one API query. However, selecting all the queries can still be cumbersome, especially given the clunky user interfaces. It is also possible to generate these API calls without the generator.

Organisations which offer the API in the form of a library also face a similar issue with the initial set-up of the indicators being time-consuming. The challenge arises from the need to know the dataset name beforehand to access it. For example, OWID offers a Python library, in beta mode, featuring a search function. However, users will still need to look through the available datasets to decide which one they want to track. Similarly, in HDX, users must browse and specify the datasets they intend to retrieve data from.

Implication on design: Given every organisation has a unique API structure, the design of the prototype will require creating individual API call formats for each organisation. There cannot be a standardised structure that can be reused for all organisations. For the prototype, a predetermined set of indicators will be selected in advance. If no generator exists, the indicators will need to be manually identified and called.

3.3.1.4 Issues with APIs: Documentation and Accurate Information

Having usable APIs is also integral to developer productivity as it will require less searching and support and should be intuitive [Piccioni et al., 2013]. Usable in this scenario could be in terms of the documentation but also aspects such as being able

to create the API request and use the API. Once an API has been created, it is also essential to ensure it is maintained. Maintenance of an API is a time-consuming task and consequently, can easily become out of date [Subramanian et al., 2014]. Unambiguous API documentation has also been found to be critical for the efficient development of software. API documentation lacking aspects such as usage examples causes productivity issues for developers and providing these would reduce errors and improve the success rate of being able to use their API [Sohan et al., 2017][Robillard and DeLine, 2010].

There is also the challenge of information on the website being outdated. This normally happens when an API is being replaced or moved to a different platform. For example, UNESCO's UIS API was taken down in 2020 with an anticipated replacement in 2021. However, as of November 2023, the same message remained. After emailing UNESCO, they replied to let me know the API would be ready for February 2024. In February 2024, the API had still not been published and after checking in, they replied the API would be ready for April 2024. This page has now been taken down and redirects to their Bulk Data Download page. Before this redirection, the only information available was the message saying the API would be released in 2021. The specific details regarding the postponed release dates were only obtained through direct inquiry via email. Information not being updated can lead to difficulty in planning for the development of projects and also a loss of trust when the information given is unreliable.

Implication on design: For the prototype, I will use data sources which have published and functioning APIs.

3.3.1.5 Issues with APIs: Dependency and Reliability Challenges

Another challenge regarding APIs is the dependency on the organisation to keep the API stable and running. APIs can experience downtime and when they go down, there's little users can do to mitigate this challenge. Furthermore, as mentioned in the previous section, they can also be completely taken down while it's being moved or replaced leaving a broken link to that organisation and a potential loss of functionality. Additionally, when the API is taken down, it's not known exactly when an updated/fixed version will be released.

Implication on design: To account for APIs going down, the prototype will need to have a log page displaying the status of all the APIs connected to the prototype. This would allow the end user to know the data from that organisation is not being updated.

3.3.1.6 Issues with APIs: Maintainability

API updates are usually explained on their website when taken down, which requires us to update our scripts to match the new API structure. For example, in Figure 3.3, there is an example of a change made to the SDMX API for ILO. This type of change would require the data structure mapping to be edited and without this change, the script would break for this query. Therefore, strong error handling is also required.

Implication on design: In the unlikely case that an API is updated when the two-week

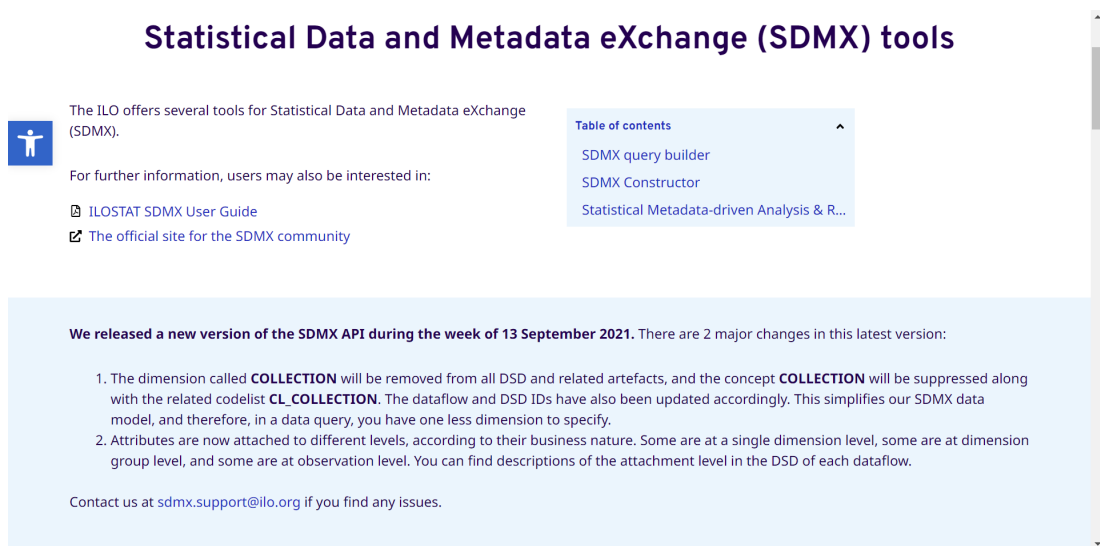


Figure 3.3: Example of a notification of changes to ILO'S SDMX API

user study is being carried out, I will treat the API as being down due to time constraints of editing the structure and getting the API connection back up.

3.3.2 Web Scraping

Finally, for the data sources that did not offer an API or library, I looked into the feasibility of web scraping them.

3.3.2.1 Issues with Web Scraping: Varied Layout and Maintainability

The websites have different layouts so if we were to web scrape, the script would have to be tailored to each organisation's page. Furthermore, if the website is updated, the existing script would be ineffective. This would then require a lot of maintenance, requiring the management of multiple scripts for multiple websites.

Implication on design: There would need to be a separate web scraping script for each website and error handling and logging mechanisms to be able to flag layout changes, enabling alerts to prompt script adjustments. Thus, this project will primarily focus on linking websites with APIs due to the time-consuming nature of web scraping.

3.3.2.2 Issues with Web Scraping: JavaScript-heavy Content

Some organisation's websites are dynamically generated and use a lot of JavaScript to render their content. This can make it difficult for web scraping tools to extract data from them. For instance, Save the Children's website has a world atlas feature that is JavaScript-based and requires user interaction, making it harder to scrape information.

Implication on design: This complicates the data extraction process, especially when dealing with interactive features. Even with browser automation, it would not be feasible to consistently web scrape and maintain the web scraping script. This makes it difficult

to efficiently extract data from these websites and so, this prototype will not include sources from such websites.

3.3.2.3 Issues with Web Scraping: Dependency and Reliability Challenges

Similar to the possibility of APIs being down, web scraping also relies on external factors such as network connectivity, server availability and the website's up-time. Any problems with the website's uptime can lead to issues when retrieving data.

Implication on design: Similar to the implications due to API network problems, it will be important to have a log of the status of the website connection to keep the user updated. Due to time constraints, my primary focus will be on linking APIs and addressing API network issues first.

3.3.3 Downloadable Resources

Some organisations offer downloadable resources. For example, UNESCO and ILO-STAT offer a bulk data download service which provides large datasets for different information that you can download. The bulk data download files are too large to open using something like Excel and need to be processed first. However, UNESCO also offers tutorials on getting started. Some organizations provide smaller datasets in downloadable formats like Excel spreadsheets.

3.3.3.1 Issues with Downloadable Resources: Automating Challenges

Downloadable resources are better suited to be used for one-time data analysis rather than being downloaded and checked regularly due to their size. It is not efficient to download and upload it into the database daily, as it would consume a lot of time and resources. Although it is possible to web scrape and automate the download from a website, any changes made to the website's layout could cause issues in finding the file to download and result in a broken link to the organization.

Implication on design: Each website would need a custom script to automate the download of the resources for each organisation. Again, for this project, I will primarily focus on linking API's due to time constraints regarding automating the process of downloading and processing it.

3.3.4 Global Data Extraction Feasibility Overview

3.3.4.1 Overview

I judged all the data sources on how easy it is to access the data based on these factors:

- API availability
- Web scraping viability (if there was no API provided)
- Response of the data format

Overall, due to time constraints, I have decided to focus on implementing the organisations which offer an API because it is the most effective and efficient method to extract data. To understand the challenges around live-linking APIs, I have chosen 5 organisations to live link: UNICEF, World Bank, World Health Organisation (WHO), HDX and Armed Conflict Location and Event Data Project (ACLED). I've chosen these five in particular because they all offer a slightly different way of linking to their API. By covering various techniques of extracting data from different APIs, I can investigate the difficulties involved in handling various types of API calls. Additionally, it enables me to test all the different types of APIs that are available in the list, thus providing me with an overview of the challenges in linking those with available APIs without having to go through all sources that have APIs available.

3.4 Prototype Data Sources

Within the data sources which offer an API, I've decided to focus on API sources covering a large number of factors and also different methods of connecting to the API.

The sources which cover the most number of factors and had the best ability to extract data are UNICEF and The World Bank. The largest data source is UNESCO however, unfortunately, since UNESCO does not offer an API, I decided to leave UNESCO out of this prototype.

The second largest data source is UNICEF which offers a SDMX API and also a website to generate the API query easily. The World Bank offers a Python library which carries out the API calls. WHO offers a similar API but not as a library. HDX offers a Python library and downloads the dataset with the API. Finally, ACLED also uses API calls but requires authentication and has a different response layout.

3.4.1 Indicators for prototype

Due to limitations in the storage capacity and API request limits of Google Sheets, I have preselected indicators for the prototype in collaboration with my user study's participant. The participant suggested finding indicators to do with out-of-school rates, enrollment, attendance and expenditure because these factors normally have discrepancies or are estimated across organisations. Furthermore, for ACLED we have decided to focus on "Violence against Civilians" as an area to get data from to lessen the amount of data being added to sheets. For WHO data, the focus is on indicators associated with the keyword "stunting". This allows the participant to then select specific indicators from this list for the user study, allowing the user study to be closer to a real-life scenario.

3.4.2 Python Data Format

The data obtained from the data sources will be transformed into a specific structure - a list of years followed by a list of values associated with those years for each indicator. This structure will be added to Google Sheets with each list as separate columns. The header will consist of the title of the indicator along with the Year for the years column,

while only the title of the indicator will be used for the values column. You can find an example of a spreadsheet in Google Sheets in Figure 3.4.

It is important to note that Google Sheets has limited functionality as a database and cannot carry out complex data manipulations or maintain relationships between sheets. Therefore, I have used a basic data format to make it easier to upload the data into Google Sheets with each sheet representing a different organization. The data is stored in columns instead of rows because it will eventually be pulled into Power BI, and Power BI's transform data feature supports transformations in this format. It also has features such as converting the top row as a header which would not work if the data was formatted with the data in rows.

	A	B
1	Adjusted net enrollment rate, primary (% of primary school age children) Year HDX	Adjusted net enrollment rate, primary (% of primary school age children) HDX
2	1993	26.766
3	1974	26.82221
4		
5		
6		

Figure 3.4: An example of the data format in Google Sheets

3.5 Power BI Design Implications

Firstly, Power BI on the free tier has limited functionality and cannot publish the report from Power BI Desktop to the Power BI Service on the web - which is an essential part of this project [pow, Accessed 2024]. Microsoft offers a 60-day free trial of Microsoft Fabric which allows you to trial as a Power BI Pro account. This allows access to features such as auto-refreshing the dashboard at scheduled times and being able to publish the report to the Power BI Service [Microsoft Fabric].

Power BI offers a REST API but it is limited to managing content and some admin operations [PowerBI Microsoft Learn]. It does not allow us to use features such as the "Get Data" feature from their API so this could not be an option to automate. Therefore, this has to be done in the interface during the set-up of this system. Microsoft Power Query function aims to support data connections from various sources including those with built-in connectors, generic interfaces such as REST API's and even website data extraction [Microsoft Power Query, Accessed 2024]. However, after an initial trial of this program, the feature did not function with some of the organisation's APIs and also did not extract data very well from those without an API. Furthermore, Power BI's API limitation also prevents the automation of the set up of these data sources through a script.

An alternative application is Microsoft Power Automate, which is a process automation platform [Microsoft Power Automate, Accessed 2024]. This is a paid solution however, even with this, linking all the APIs may not be feasible because some respond with file downloads rather than a JSON response.

After experimenting, I discovered that connecting to data sources and Power BI directly was not feasible. Therefore, I opted for a database connection instead. Power

BI offers connections to a wide range of databases [Power BI, Accessed 2024a] however since the end Power BI system would be online (the published Power BI report), the database used would also need to be online. Using an on-prem (local) database would require using a gateway server to link the database to Power BI [Power BI, Accessed 2024c]. However, due to a lack of access to such servers due to them all having a charge associated with them, this option was not feasible for the initial prototype. This resulted in the options for this prototype being databases on the cloud. Major cloud service providers provide suitable solutions however again require a cost to run, especially when storing large amounts of data. Thus, an alternative was to use a cloud spreadsheet which is already online and can be accessed from Power BI. Hence, Google Sheets was chosen as the database of choice for this prototype.

Additionally, through interacting with Power BI, I've found that every time a change is made e.g. a column is deleted or added to a table or a new data source is added, it needs to be re-published to apply the changes. Issues with adding/removing columns and Power BI detecting this means being able to add new indicators or new data sources will not be possible.

Overall, this leads to the following workarounds. There will need to be an initial set-up on Power BI Desktop to connect Power BI to the data source due to Power BI's API limitation. The database it will be connected to will be Google Sheets. The initial prototype will not have any features to add/delete indicators due to Power BI's limitation.

3.6 Chapter Summary

This chapter covered the design of the prototype, which included the requirements of the prototype and system overview. The chapter also discusses data extraction methods provided by global data sources and details the complexities associated with these methods.

Chapter 4

Implementation

4.1 Prototype Implementation

4.1.1 General Process

The general process for implementing the connection to the five chosen organisations has been split into two parts, acquiring data for the chosen indicators into Google Sheets and getting data from Google Sheets into Power BI.

API data to Google Sheets process (Python script):

1. Connect to the API with Python
2. Extract data for the chosen indicators from the API.
3. Transform and clean data to the correct format
4. Load into Google Sheets

Google Sheets to Power BI process (carried out manually):

1. Connect Google Sheets to Power BI
2. Extract data from Google Sheets
3. Transform data types in Power BI
4. Load into Power BI Model
5. Publish Power BI Report

The Google Sheets to Power BI process was mainly the same for all organisations but could not be scripted and automated due to Power BI's API limitations. However, since this connection only needs to be made once, it is not a significant issue. The data cleaning process was also the same for all of the organisations. This involved removing null and non-numeric values. The API data to Google Sheets process varied as the organisations had different APIs.

4.1.2 API data to Google Sheets Processes

This process was carried out using Python. There was a separate script for each of the chosen organisations and a main script for setting up the indicators and running the scripts. This page would be the script which would be run daily and link all the different calls for the scripts together.

4.1.2.1 UNICEF

The implementation required generating an API call using the UNICEF Indicator Data Warehouse. The UNICEF data response was then cleaned and transformed to the format decided for inserting into Google Sheets, and then loaded into UNICEF sheet through a Python script custom-made for the UNICEF response.

Example of UNICEF's API call and response:

Example Indicators: Government expenditure on education (% GDP) and Adjusted net attendance rate for children of primary school age

HTTP API Request Call: `https://sdmx.data.unicef.org/ws/public/sdmxapi/rest/data/UNICEF.AFGHANISTAN_CO,AFG_CO,1.0/AFG.ECON_GVT_EDU_EXP_PTGDP+ED_ANAR_L1...?format=sdmx-json`

JSON Response:

A condensed representation of the JSON response from this API call is shown in Figure 4.1. While UNICEF provides documentation for the structure of their API response, it is slightly outdated and not fully accurate [UNICEF, Accessed 2024a]. However, it is still possible to interpret the structure from the documentation with some analysis of the data.

The JSON response contains a hierarchical structure with several key-value pairs. Within the response, there is a key “dataset” which holds the main data. Inside “dataset”, there is a key “series” which contains the values for each indicator. These indicators are uniquely identified by a specific format consisting of five numbers separated by colons. e.g region:factor:sex:age:subnational-level

Also within each “series”, there is the key “observations” which hold the value for that indicator and details such as the year.

There is also another key “structure” which holds information about the structure. Inside “structure”, there is a key “dimensions” and inside “dimensions”, there is another key “series” which holds information regarding each attribute and which position it links to.

For instance, the identifier 0:1:1:0:0 refers to Afghanistan's Adjusted Net Attendance Rate for Children of Primary School Age, Female, Total Age, Administrative Level 0. The first number (0) corresponds to the 0th value in the “values” section of the “structure's” series. The second number (1) corresponds to the indicator “Adjusted Net Attendance Rate for Children of Primary School Age”, which is the 1st value in the

”values” section of the ”structure’s” series, and so on. Additionally, there is a key called ”observations” that contains information about different time periods.

```

1- {
2-   "meta": {...},
3-   "data": {"dataset": [
4-     {"links": [...],
5-      "action": "...",
6-      "series": { "0:0:0:0:0:0": {"attributes": [], "observations": {"0": ["4.058869839", null, null, null, 0]}, ...},
7-                  { "0:1:1:0:0:0": {"attributes": [], "observations": {"0": ["53.200001", null, null, null, 1]}, ...}}}
8-   ]},
9-   "structure": {
10-     ...
11-     "dimensions": {
12-       ...
13-       "series": [
14-         { "id": "REF_AREA", "name": "Geographic area", "keyPosition": 0, "role": null,
15-           "values": [{"id": "AFG", "name": "Afghanistan"}]},
16-         { "id": "INDICATOR", "name": "Indicator", "keyPosition": 1, "role": null,
17-           "values": [
18-             { "id": "ECON_GVT_EDU_EXP_PTGDP", "name": "government expenditure on education (% GDP)" },
19-             { "id": "ED_ANAR_L1", "name": "Adjusted net attendance rate for children of primary school age",
20-               "description": "Percentage of students of the official primary school age group who attended primary or secondary
21-                 education at any time during the reference academic year"
22-             }
23-           ]
24-         }, ...]
25-       }
26-     }
27-   }
28- }

```

Figure 4.1: Example of the JSON response from UNICEF’s API call

4.1.2.2 World Bank

This required manually finding the indicator codes for each of the chosen indicators from their Open Data website. After obtaining these, the API calls would be made using the World Bank’s Python library and the data would be cleaned and transformed to the format required for Google Sheets, and then loaded into a World Bank sheet.

Example of World Bank’s API call and response:

API Request Call: Using the World Bank’s library was fairly straightforward. It required getting the indicator codes and names into a dictionary and using that when fetching data for those indicators from the library.

Response: The API library returned a pandas data frame in a table format with the columns as years as rows as indicators which is fairly easy to transform.

4.1.2.3 ACLED

This required creating an account, getting an access key from ACLED and using this in the API calls. The ACLED response was then cleaned and transformed into the format needed for inserting it into Google Sheets, and then loaded to an ACLED sheet.

Example of ACLED’s API call and response:

HTTP API Request Call:

```
https://api.acleddata.com/acled/read/?key={self.api_key}&email={self.email}&country={country}
```

JSON Response: The main information from the response can be obtained in the “data” section of the response where each dictionary is an event. Inside each event dictionary,

it contains information about the event date, the type of event and location and notes. A shortened version of the JSON response can be seen in Figure 4.2.

```

1 {
2   "status": 200,
3   "success": true,
4   "last_update": 1,
5   "count": 5000,
6   "messages": [],
7   "data": [
8     {
9       ...
10      "event_date": "2024-03-08",
11      ...
12      "disorder_type": "Political violence",
13      ...
14      "country": "Afghanistan",
15      "admin1": "Herat",
16      "notes": "...", ...},
17     ...], ...
18  }

```

Figure 4.2: Example of the JSON response from ACLED's API call

4.1.2.4 WHO

WHO's API offers a method to get indicators which contain a key term. This was used to get the indicator codes for the chosen indicator code. The response for the data from the API call to the indicator codes was then cleaned and transformed to the correct format required for Google Sheets, and then loaded to a WHO sheet.

Example of WHO's API call and response:

HTTP API Request Call: [https://ghoapi.azureedge.net/api/{indicator_code}\\$filter=SpatialDim%20eq%20%27{self.country_code}%27](https://ghoapi.azureedge.net/api/{indicator_code}$filter=SpatialDim%20eq%20%27{self.country_code}%27)

JSON Response: WHO's API response is a list of dictionaries inside "values" with each value as a separate dictionary. Inside each values dictionary, there is information regarding the year, the value and any comments.

```

1 {
2   ...
3   "value": [
4     {
5       Id: 4004320,
6       IndicatorCode: "NUTRITION_ANT_HAZ_NE2",
7       SpatialDimType: "COUNTRY",
8       SpatialDim: "AFG",
9       ...
10      TimeDim: 1997,
11      ...
12      Value: "53.2",
13      ...
14      Comments: "Converted estimate Age 0-5 months not covered; Unadjusted for age",
15      Date: "2022-06-16T09:28:10+02:00", ...
16    }, ...
17  ]
18 }

```

Figure 4.3: Example of the JSON response from WHO's API call

4.1.2.5 HDX

HDX also offered a Python library which gets desired data given the name of the dataset. HDX's response was a CSV file with the data so this required reading from the CSV file, cleaning and transforming to the correct format for Google Sheets and then loading to a HDX sheet. The downloaded CSV files were then deleted.

Example of HDX's API call and response:

API Request Call: Using HDX's library was also fairly straightforward. It offered a download method to download the chosen dataset.

Response: The response were CSV files - the main csv file having the information in the columns country, country code, year, indicator name, indicator code and value. There are sometimes multiple csv files depending on the resources and this could contain other information e.g. quick chart indicators.

4.2 Software Testing

To ensure the prototype is robust, I've carried out unit and integration tests. This was challenging to test because the code uses a lot of external APIs so the code coverage is lower for the scripts where there are more sections to do with the external API calls. For example, HDX's script involved many external links and thus has a much lower code coverage, however for the purpose of the user study, this will still work. I've used PyCharm's code coverage to generate a coverage report [PyCharm Code Coverage]. The code coverage for the scripts relating to each of the organisations can be found in Figure 4.4. I've used Python's unit test library to test each organisation's scripts. I've used mocking to mock the external APIs and created tests to cover different scenarios (successful and failed data fetching), tested the data transformation for each organisation and edge cases such as datasets with missing values.

organisation	100% files, 82% lines covered
ACLED.py	81% lines covered
HDX.py	40% lines covered
UNICEF.py	93% lines covered
WHO.py	97% lines covered
WorldBank.py	100% lines covered

Figure 4.4: Unit and integration tests for each of the organisation's scripts

Another challenge was mocking different formats of data. For example, UNICEF, WHO and ACLED gave JSON responses as they were HTTP API calls. However, World Bank and HDX used Python libraries and HDX's response was in a CSV format. Thus, this required different styles of tests and different response objects were mocked e.g. mocking the JSON response and mocking a CSV file being generated and read.

4.3 Requirements Testing

The following table looks into whether the system prototype meets the requirements set out during the design of the project through testing. I will then look further into any requirements that are not fully met to understand why.

No.	Requirement	Met?
1	The system should connect multiple global web data sources to the Power BI system	Partially
2	The system should show the latest updates (new data records and updates to existing records) from each of the data sources (daily).	Met
3	The system should make it clear what has been updated when a data source has been updated.	Partially
4	The system should make clear when the data was last checked for updates.	Met
5	The system should display the connectivity status to the different data sources.	Met
6	The system should be accessible from anywhere with an internet connection.	Met
7	The system should be robust. It should be able to handle issues such API's timing out, API's returning incorrect responses.	Met

Requirement 1

This requirement is met partially as it can connect multiple global web data sources to Power BI however it is limited to ones with working APIs. Not all websites are web scrape-able and those that are would require highly customised scripts. The process to connect APIs is already fairly customised but the process to customise web scraping for multiple different websites would not be practical.

Requirement 3

This requirement involves the need to identify which specific data points within an indicator have been updated, and to display these updates in an evaluation. The current prototype, which utilizes Google Sheets and Python, makes it difficult to detect these updates since the entire sheet would need to be looped through. However, with a structured database, this process would be easier to perform.

4.4 Chapter Summary

This chapter discusses the prototype implementation, including API call responses and requirements met. It also covers software testing to prepare for user testing.

Chapter 5

Evaluation

5.1 User Study Design

I have decided to evaluate the system by conducting a user study. Since the system needs to be tested over a longer time period, theoretical evaluation methods were unsuitable. As a result, I designed a user study in which the end user tests the product for two weeks. The study involves a mid-study semi-structured interview at the end of the first week, followed by a final semi-structured interview at the end of the second week.

5.2 End Users

5.2.1 General End Users

The data sources the system will be working with are all publicly available datasets and these are currently used by a range of users. These users include Humanitarian Aid Workers, Researchers and Academics, Policy Makers, other NGOs and non-profits, journalists and more. This project will also have similar end users.

5.2.2 User Study End User

The user study will be carried out on one participant, an academic who would use a system such as this for their research. The participant is involved in the parent project and is also one of my supervisors. Given the constraints of the user study timeframe being two weeks and the nature of the project, the chosen participant serves as an ideal candidate who is familiar with the project and would also be an end-user of such a system. Although the participant is one of my supervisors and involved in the parent project, they are not directly involved in the technical aspects of the project, allowing them to give insights and feedback regarding the functionality and relevance of the end system. Academics would use a system such as this to allow for more comprehensive and timely analysis and to be able to carry out comparisons between organisations. Thus, as an academic, they will be able to give valuable insights into the usefulness of such a system for their work and provide valuable feedback.

In this scenario, a user study on one user is reasonable for this stage of the project as the study is mainly evaluating the feasibility and functionality of the system and its effectiveness. The user population is also niche so it would be challenging to recruit multiple participants from the list of end target users. Given the nature of the project and the system needing to be used for a longer time frame to fully get a sense of using the product, a single-user study is sufficient to gain feedback and insights into issues for further improvements for now.

5.2.3 NASA-TLX Evaluation

As part of the evaluation of the proposed system that aims to reduce attributes such as performance, effort and time taken, the participant will complete a NASA-TLX assessment. This is a subjective workload assessment tool that allows us to measure the operator's workload at different stages of the system's operation [Hart and Staveland, 1988]. It has been widely used in different areas of a system, including interface design, automation, and model validation [Hart, 2006].

For this project, NASA-TLX will be used to evaluate the impact of the automated system of live linking global data sources to Power BI. The participant will rate each of the factors on a scale, and then decide which factors in the pairs from the given comparison cards were more significant for the task. This will then determine the weighting given to the factor.

To make the evaluation more effective, a slightly modified version of NASA-TLX will be used. Instead of filling out the form after every main task, the participant will only complete it at the end of the study. This will allow us to observe the full effect of the system after a longer period of use. Thus, the participant will complete the NASA-TLX at the end of the second interview of the user study.

5.2.4 User Study Process

Several key methodologies and tools have been utilised in conducting the user study to evaluate the system's functionality and user experience.

The data will be linked along with logging tables showing what indicators had been changed, when the data was last refreshed and the connection status to the different organisations. The participant will have access to this data and will be able to create reports relevant to them on Power BI. Using a semi-structured interview provides us with the flexibility to explore any issues that may arise during the interview process. Conducting both a mid-study interview and a final interview will ensure that the study is progressing smoothly and help identify any significant changes that need to be made. As part of the evaluation process, the participant will fill out a NASA-TLX form to provide feedback on aspects such as effort, frustration and performance. This gives insight into the system's improvement as it saves users time and frustration of going through individual data sources.

Furthermore, the data was updated very sparsely and unpredictably which meant the probability of catching updates in two weeks was considered remote. Therefore,

the decision was taken to make some changes manually in the database. This would also compliment any changes that were real in the short time frame and thus, should not affect the overall representation of the system. Furthermore, the participant was not told when updates would be made and so there was still an element of uncertainty which is accurate to real life.

5.2.5 Materials

The materials used were a Participant's Information Sheet (see Appendix A), a Participant's Consent Form (see Appendix B), the forms to carry out the NASA TLX (see Appendix C) and the list of semi-structured interview questions (see Appendix D).

5.3 User Study Results

5.3.1 User Study Set Up

The user study began with an initial call to configure access to the published Power BI Service. The participant created one graph for each selected data source's indicators. These indicators were selected as they resemble indicators they would use while working. The chosen indicators for each of the data sources for this study can be found in Table 5.1.

Table 5.1: Chosen indicators for each data source

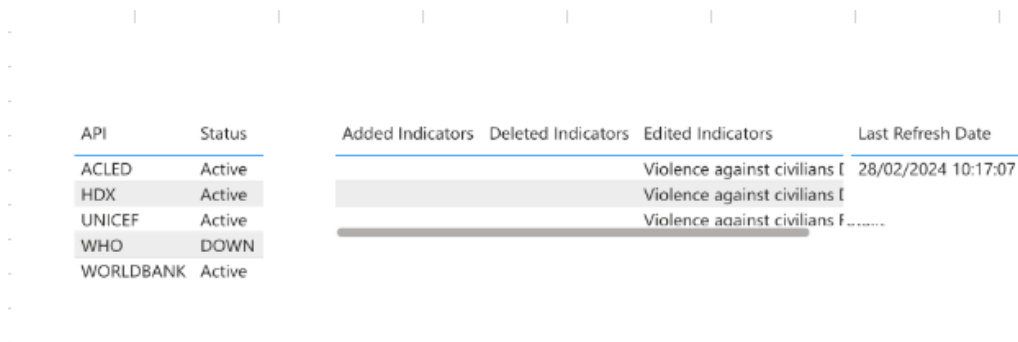
Organisation	Indicator
UNICEF	Afghanistan Completion rate for children of primary school age total
ACLED	Violence against civilians fatalities
World Bank	Government expenditure on education, total (% of GDP)
HDX	Number of deaths ages 5-9 years
WHO	Stunting prevalence among children under 5 years of age (% of height-for-age <-2 SD) survey-based estimates

The participant kept screenshots of the original graphs and the log tables. They then checked for updates twice a day, once in the morning and once in the evening, and noted changes in the indicator along with any relevant screenshots in a spreadsheet. I could also monitor this spreadsheet to monitor any changes. The mid-study semi-structured interview took place on the following Monday and the final semi-structured interview took place on the Monday after the end of the study. The main questions for the mid-study and final semi-interview can be found in Appendix D. These questions focused on the functionality of the system, the value of the system and the issues faced.

5.3.2 Mid-study Semi-structured Interview Findings

During the semi-structured interview, the participant shared their experience of using the system and provided feedback on what worked well, what could be improved, and any additional comments they had. The main comments can be broken down into the following themes:

Positive feedback and reassurance: The participant said they found the system “reassuring” and thought it would “save masses of time”. They mentioned they “liked the setup page like the logging page” (see Figure 5.1).



The screenshot shows a table with two main sections. The left section lists APIs and their status, and the right section shows a log of indicator updates.

API	Status	Added Indicators	Deleted Indicators	Edited Indicators	Last Refresh Date
ACLED	Active			Violence against civilians [28/02/2024 10:17:07
HDX	Active			Violence against civilians [
UNICEF	Active			Violence against civilians F.....	
WHO	DOWN				
WORLDBANK	Active				

Figure 5.1: The participant's screenshot of the logging tables they had created showing the indicators which have been updated and WHO API being down

Need for change tracking feature: They mentioned knowing what has changed has been useful however they suggested being able to see historical data would be better. The participant highlighted the importance of needing to know what exactly had changed within the indicator and what the data was before, especially when dealing with large datasets.

Issues: The participant noted a problem where the order of years was disrupted despite setting it to be ascending. This occurred when a change was made, which was a result of the structure of the prototype's data in Google Sheets. To address this issue effectively, a check should be incorporated to prevent unintentional changes to the data order, ensuring consistency for end-users. Additionally, they mentioned their inability to create comparison graphs but expressed the usefulness of having that ability. The full list of issues can be found in Table 5.2.

5.3.3 Mid-study Changes Findings

We went through the daily logs of the participant and checked if they noticed the changes made. The participant successfully noticed both changes made that week regarding UNICEF's data being added and deleted. The participant also noticed the real ACLED data coming through. The data changes made can be found in Table 5.3 and the real changes coming through from the live link can be found in Table 5.4.

Table 5.2: List of Issues and Comments from Week One of the User Study

Date	Issue	Comment
22/2/24	Unable to save report on iPad	Limitation of Power BI
26/2/24	Unable to create comparison graphs	Limitations of Power BI and Prototype
26/2/24	Unable to edit report once it was saved	Limitation of Power BI
28/2/24	Refresh confusion	This was resolved in mid-study interview
29/2/24	WHO Stunting data states its been updated but no updates	Result of WHO being down and API changes changing the structure of the table

Table 5.3: Week 1: Data Changes I made

Date Changed	Organisation	Data Changed	Date Noticed
28/2/24	UNICEF	Added data for 2024's Completion Rate for Children of Primary School Age Total	29/2/24
1/3/24	UNICEF	Removed the data for 2024's Completion Rate for Children of Primary School Age Total	4/3/24

5.3.4 Post-mid-study modifications

Following the mid-study phase, modifications were made to tackle the limitation of not being able to generate comparison graphs, which was discovered during the initial week of the study. This modification was crucial for the study because it allowed for a more comprehensive analysis by being able to compare indicators from various organisations within a single visualisation. Furthermore, this provided a more authentic experience for the participant as it aligned with their real-life needs and usage scenarios as an academic.

The data was then moved into one sheet to allow data from different organisations to be put on one graph. In the previous set-up with the organisations on different sheets, Power BI did not allow graphs with data from different sheets on the same graph. The issue regarding not being able to make graphs was that the data was spread across multiple tables and Power BI did not allow graphs to be created with data from different tables. While it might not be practical to put all the data into a single table, the data is stored in one table during the second week to allow us to test and create comparison graphs.

Table 5.4: Week 1: Live Linked Data Changes

Date	Organisation	Data Changed	Date Noticed
27/3/24	ACLED	Added data for Violence Against Civilians Fatalities	28/2/24
27/3/24	WHO	API was down	28/2/24

The participant created 4 more comparison graphs, comparing indicators from different organisations in the same graph. The indicators chosen to be tracked by the participant can be found in Table 5.5. The participant continued the same process as the previous week, checking both sets of graphs twice a day for changes and noting down changes noticed and any issues along with screenshots.

Table 5.5: Indicators chosen to be compared by participant

Organisations	Indicators being compared
UNICEF and World Bank	Afghanistan Government Expenditure on Education (% GDP) Total
UNICEF and World Bank	Afghanistan Out-of-school rate for children of primary school age (Male and Female)
UNICEF and World Bank	Afghanistan Primary School Completion Rate
UNICEF	Afghanistan Out-of-school rate for children of primary school age, Afghanistan Primary School Completion Rate, Afghanistan Adjusted net attendance rate for children of primary school age

5.3.5 Post-Study Semi-Structured Interview Findings

The main findings from the post-study semi-structured interview can be broken down into the following themes:

Appreciation for comparison graphs: The participant appreciated the system's ability to create comparison graphs and said the feature felt more "real" and is the type of thing they would "want to be able to build in", stating it "gives you more opportunity to see the changes more easily".

Significance of including years: They mentioned the significance of having the years included, stating it provides an audit trail, which increases confidence in the data. In the system, users are able to access data from previous years in addition to the most recent data available.

Issues: Initially, there were a few issues due to the refresh date being wrong as a result of the modifications made. However after this was addressed, the rest of the week's data processing ran smoothly.

There was also some confusion surrounding the timing of refreshes. The Power BI application's refresh schedule feature was set to certain times and Power BI would read from the Google Sheets at those times. However, the script only ran once a day to add data to the Google Sheets. It was clarified that refreshes update the connection to the database, but the database itself undergoes a refresh only once daily. Therefore, it is important to ensure that both the refresh and the database are synchronized for accurate data management.

5.3.6 Post-Study Changes Findings

Table 5.6: Week 2: Data Changes I made

Date	Organisation	Data Changed	Date Noticed
6/3/24	World Bank	Added data for 2023's Out of School Data	7/3/24
7/3/24	UNICEF	Added data for 2023's Government Expenditure on Education Total	8/3/24
7/3/24	World Bank	Edited data for 2009's Government Expenditure on Education Total	8/3/24

The participant successfully noticed the changes made to the World Bank's out-of-school data, UNICEF's expenditure data and the World Bank's expenditure data. The date the changes are noticed is a day after they were made. This is due to the way the prototype was scheduled to run. The data in the database was modified the previous day but the system retrieves data at 6 am, so changes made during the previous day would only be reflected the following day. The alterations made and the observed changes can be found in Table 5.6.

5.3.7 NASA-TLX Findings

For this form, the participant was tasked to evaluate the process of generating graphs and checking for daily updates in the report. The overall score was a low score of 4.67. Scores under 9 are considered low and indicate a low workload overall [Prabaswari et al., 2019]. This was only carried out on one participant but is still valuable to indicate how much easier the system has made it to carry out tasks which would be tedious/ not practical to carry out in real life.

5.3.8 Overview of User Study

Overall, the user study provided valuable insights into the functionality of the system and proved that a live link was possible, especially with the real-life ACLED data coming through. The study also confirmed the potential improvements it would bring in terms of workload which was validated through the NASA TLX assessment.

Furthermore, the participant emphasised the value of the system, noting its potential to save time and reassure users. They appreciated the ability to consolidate information from various sources into one visualisation, especially from an academic standpoint. Additionally, features such as a log page for tracking changes and downtime were deemed important.

During the study, we were able to identify some areas that require further development. These areas include the need to track the exact changes made and the history of these changes. We found creating meaningful graphs on Power BI can be challenging, and there are areas where the live-link data and Power BI can be presented better to provide a smoother user experience.

5.4 Chapter Summary

This chapter discussed the methodology and results of the user study.

Chapter 6

Conclusion

This chapter will look into discussions, lessons learned, limitations, and future work from the study's findings and the project overall.

6.1 Discussions

This section will delve into two discussions arising from the study: the challenges involved in live-linking global data sources and working with Power BI. I will then reflect on the key takeaways and lessons learned from the project.

6.1.1 Live Linking Discussion

The process of establishing live links to global data sources faces several challenges. Firstly, many websites lack API's and web scraping is impractical in most cases. Furthermore, maintaining an API requires significant customisation with each update. Although providing a full live link to all global data sources may not be feasible, those offering an API can be linked.

Furthermore, there is a strong dependency on organisations for updates and notifications regarding API changes. A lack of communication from organisations can lead to inconveniences, as experienced with the World Bank API going down during testing.

Additionally, outdated information on webpages further complicates the process, as experienced during the early stages of the project with UNESCO having outdated information regarding their API release date. This dependency on API status can cause delays, as seen in this project. However, organisations have been responsive when contacted directly from my experiences of emailing UNESCO and the World Bank regarding information about their API being down. Although, their information was not available on their public websites.

Overall, navigating these challenges highlights the need for improved communication and accessibility of API information to simplify the live-linking process.

6.1.2 Power BI Discussion

The process of setting up can be cumbersome, as it requires customization and integration with end-users from the beginning. This is necessary to determine the specific indicators to be used or whether they prefer all the indicators.

Furthermore, the Power BI limitations with needing to use the Power BI desktop and republish every time a new data source is added is not practical for the end solution as it cannot be automated. As a result, anything to do with adding and deleting data sources needs to be done manually and thus a solution to this needs to be identified.

Although Power BI is popular for its ease of use among non-technical users, it can be difficult to create more complex graphs with lots of data. This was identified during the user study however it might also be due to the unconventional layout used in our user study.

However, it was quite straightforward to upload the data to the cloud, which made it easily accessible to everyone through the publish feature. The only drawback is that to access all features, such as freely creating custom reports, the user needs to be a member of the same organization as the original user who has published the Power BI report.

6.1.3 Lessons Learnt

The main lessons learnt regarding how organisations should share and keep their data to make it more accessible are as follows:

Standardised API call and response structure: Having a standardised method of calling APIs and following a set response structure would be useful for building a system such as a live-linked platform and for developers in general. It would save the time needed to understand how the API works and what the response structure is like and would allow for easier automation of data from all of these types of organisations as one script would work for lots of organisations' data. The SDMX is an example of such an initiative however currently is only being used by some organisations as found during the design stage. An example of an improvement would be for more organisations to adopt a standard such as SDMX to make the process of dealing with statistical data easier. This would also solve the issue of standardising the way to get all the desired indicators in the same way without having to manually find them in the metadata.

Up-to-date information: It is important to ensure the information regarding API availability is accurate and up to date. The consequences of this can significantly hinder development in applications wanting to use their APIs, especially if something is planned to match the date a particular API is released. Furthermore, when there have been changes, it is important to ensure the documentation is also updated to reflect these changes to make it easier for developers to use. Additionally, it would be useful for developers to have a page to check if particular APIs are down and have an easy-to-access contact email for the team in charge of APIs e.g. Mastercard has a status page for their APIs with a contact email also listed [Mastercard, Accessed 2024].

Effective versioning practices Organizations often take down their older API when

updating it, leaving developers with no access to it during the update process. A better approach would be to keep the old API running until the new one is fully operational, and then replace it. This would ensure uninterrupted access to the API for developers. They could make use of versioning however since there is a lack of a centralized way of versioning in Web APIs has led to many creating their own ways to version meaning there are many different and inconsistent practices. Versioning will allow for compatibility with existing clients and also allows organisations to make updates etc. and a consistent versioning method used across these organisations will make it easier to use and automate processes [Serbout and Pautasso, 2023].

Offering an API Offering an API is the easiest way to be able to share data with others who can use the data in their own platforms. API's have been stated to be a key enabler of interoperability by the European Commision [Borgogno and Colangelo, 2019]. It can be used for all use cases e.g. analysts for one-time use and use cases such as ours for more frequent use and updates. Especially if the other factors are taken into consideration e.g. standardised format and response, it would make it very easy to use and extract data from many organisations with one script.

Overall, if organisations did this, creating a live link would be much simpler and require less effort for the developer.

6.2 Limitations

Firstly, a large limitation of the user study was only having one user in the study. As a consequence, the results obtained from the study may not be representative of the broader population, as they only reflect the perspective of one category of end-users and also only one user. However, given the niche target end-users, it was challenging to find more users from within this user base to carry out a more extensive study. To improve the reliability of the study, it would be better to have more users participate in the research and ensure that users from different backgrounds are included to cover the various types of end-users.

Furthermore, to improve the project, I would involve end-users more heavily from the design stage. This would have allowed us to understand exactly what features and functionalities they wanted to see in the final product. For example, for academics, it was important to be able to compare past data changes and analyze them across different organizations. In future projects, I would prioritize involving end-users in the design stage to better understand their needs and determine what features are most important to them.

During the user study, the view of the published Power BI the participant had access to was limited due to the set-up of Power BI. For the free trial, I created a Microsoft Entra account which was used to create the Power BI reports under the domain of an organisation I had created for this trial. Reports and dashboards can only be shared with users within the same email domains so we faced some challenges connecting the participant to the Power BI service [azu, Accessed 2024]. The participant was added as a guest user to the uploaded service and this allowed them to be able to access a limited version which was not ideal, but sufficient for the study. This resulted

in them having limited ability to view and edit dashboards compared to a normal Power BI view.

Another limitation was the issue of the years not being recognised as the same years and so the visuals were slightly incorrect. e.g. when creating comparison graphs this was a limitation of Power BI and the formatting of the prototype's data types. This is something to address in future builds.

6.3 Future work

Live linking alternative data

I reviewed the live-linking of global sources, but future work needs to be done to understand how to live-link ad hoc data sources. Following this, further work could also be done to look into websites without APIs to see if there are other methods to scrape data from websites using similar techniques to getting data from ad-hoc data sources.

Dealing the data from the live-link

Live-linking all the data sources results in a lot of data; therefore, this data must be managed properly. For this project, due to the requirement of needing to have it on the cloud, the options to use a more suitable database to store data were restricted. Using an actual database would allow the data to be stored in a more structured manner allowing more detailed information, information regarding history and metadata to be stored.

Furthermore, the indicator titles can be long and the Power BI interface is limited to how many of the titles it can show, making it difficult to know which indicator is for what and making the overall interface of Power BI quite cluttered. Power BI has the option to create folders but there is no option to automate this currently and has to be done manually which is not a feasible option when dealing with large amounts of data. Additionally, the automated graphs generated are unclear and overall not intuitive to navigate on Power BI. Future work needs to be done to improve the user experience for users dealing with large quantities of data on Power BI and how to display the information.

Future Prototype Implementation

In the project, my design choices for the prototype were limited by multiple factors such as costs related to running a server to run the gateway to be able to run the system 24/7 and also connect to databases on-prem or use a cloud database. This prototype demonstrates the possibility of live-linking API's to Power BI however in reality, without the constraints I had, the prototype could be more advanced. Along with the benefits having access to a server gives, being able to use a database rather than Google Sheets would also allow for more complex queries and data be to organised easier which could help scale the platform up, add features and be able easily deal with different parts of the data. Furthermore, looking more into options such as a No-SQL database might be interesting as it might be better for dealing with data with different structures.

6.4 Conclusion

For this project, I investigated the possibility of linking multiple data sources in real-time to Power BI, with the ultimate goal of enhancing data timeliness. The project originated from scoping out an idea from the current challenges highlighted in the parent project.

To complete this project, research was carried out into all of the global data sources provided by the parent project. Through this exploration of the global data sources, it became evident only a few sources offered suitable APIs for automated data extraction. Therefore, these were the chosen sources which were investigated in more detail.

From there, a prototype system was created to link live data from different sources with APIs in different formats to Power BI. Then, a user study was carried out with a two-week test period of a potential end-user using the system in a way that would resemble how it would be used in real life. This gave valuable feedback that the system would be useful and also suggestions for additions to the system.

While Power BI offers strong capabilities for developing static pages and importing data, its support for live linking is limited. Live linking in other ways is possible as shown in the project but can still be cumbersome and requires work to create custom scripts. Additionally, there are limitations regarding automating several Power BI processes, which result in manual interventions for certain tasks. Thus, addressing these limitations is crucial for more efficient project outcomes.

In conclusion, the project validates the hypothesis that live linking different data sources to Power BI would allow users to more effectively access relevant data. By meeting specific requirements such as providing APIs, live linking offers a promising solution to enhance data accessibility and timeliness for connecting to Power BI. The results of the user study confirm the hypothesis, highlighting the potential effectiveness of live linking in streamlining the data integration processes and improving user workflow.

Bibliography

- UNESCO Institute for Statistics API Portal. <https://apiportal.uis.unesco.org/bdds>, Accessed 2024a.
- Education Estimates. <https://education-estimates.org/>, Accessed 2024b.
- Whitepaper: Azure B2B Power BI. <https://learn.microsoft.com/en-us/power-bi/guidance/whitepaper-azure-b2b-power-bi>, Accessed 2024.
- Power bi pro. <https://powerbi.microsoft.com/en-us/power-bi-pro/>, Accessed 2024.
- SDMX Technical Standards. https://sdmx.org/?page_id=3425, Accessed 2024.
- Afghanistan Education Cluster Dashboard. Afghanistan education cluster dashboard. <https://app.powerbi.com/view?r=eyJrIjoib2NjMyOWUtOGExMS00NjZkLThjYTgtNjgwM2MzMzNmQyYmJjIiwidCI6Ijc3NDEwMTk1LTE0ZTEtNGZiOC05MDRiLWFjMTg5MjAyMzY2NyIsImMiOjh9>, Accessed 2024.
- Amazon Web Services. Amazon AWS API. <https://aws.amazon.com/what-is/api/>, Accessed 2024.
- Amit Awasthi. Business intelligence: Concepts, components, techniques and benefits. *Asian Journal of Management*, 3(4):196–203, 2012.
- Oscar Borgogno and Giuseppe Colangelo. Data sharing and interoperability: Fostering innovation and competition through APIs. *Computer Law & Security Review*, 35(5): 105314, 2019. ISSN 0267-3649. doi: <https://doi.org/10.1016/j.clsr.2019.03.008>. URL <https://www.sciencedirect.com/science/article/pii/S0267364918304503>.
- Elizabeth Buckner, Daniel Shephard, and Anne Smiley. Beyond numbers: The use and usefulness of data for education in emergencies. 2022.
- Karen M Devries, Dipak Naker, Adrienne Monteath van Dok, Claire Milligan, and Alice Shirley. Collecting data on violence against children and young people: need for a universal standard. *International health*, 8(3):159–161, 2016.
- DFID - GOV. Working effectively in conflict-affected and fragile situations. <https://www.gov.uk/government/publications/working-effectively-in-conflict-affected-and-fragile-situations>, 2010.

- Tim Furche, Georg Gottlob, Giovanni Grasso, Xiaonan Guo, Giorgio Orsi, Christian Schallhart, and Cheng Wang. Diadem: Thousands of websites to a single database. *Proc. VLDB Endow.*, 7(14):1845–1856, oct 2014. ISSN 2150-8097. doi: 10.14778/2733085.2733091. URL <https://doi.org/10.14778/2733085.2733091>.
- Daniel Glez-Peña, Anália Lourenço, Hugo López-Fernández, Miguel Reboiro-Jato, and Florentino Fdez-Riverola. Web scraping technologies in an API world, Briefings in Bioinformatics. *Briefings in Bioinformatics*, 2014.
- Child Mortality Coordination Group et al. Tracking progress towards the millennium development goals: reaching consensus on child mortality levels and trends. *Bulletin of the World Health Organization*, 84(3):225, 2006.
- Sandra G Hart. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.
- Sandra G Hart and Lowell E Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.
- Johannes Hoogeveen and Utz Pape. *Data collection in fragile states: innovations from Africa and beyond*. Springer Nature, 2020.
- Hostinger. Hostinger cron job tutorial. <https://www.hostinger.co.uk/tutorials/cron-job>, Accessed 2024.
- Humanitarian Data Exchange (HDX). HDX FAQ. <https://data.humdata.org/faq>, Accessed 2024a.
- Humanitarian Data Exchange (HDX). HDX Python API Documentation. <https://hdx-python-api.readthedocs.io/en/latest/>, Accessed 2024b.
- IBM. IBM Extract, Transform, Load (ETL). <https://www.ibm.com/topics/etl>, Accessed 2024a.
- IBM. IBM REST APIs. <https://www.ibm.com/topics/rest-apis>, Accessed 2024b.
- Tino Kreutzer, Patrick Vinck, Phuong N Pham, Aijun An, Lora Appel, Eric DeLuca, Grace Tang, Muath Alzghool, Kusum Hachhethu, Bobi Morris, et al. Improving humanitarian needs assessments through natural language processing. *IBM Journal of Research and Development*, 64(1/2):9–1, 2019.
- Anders Koed Madsen, Mikkel Flyverbom, Martin Hilbert, and Evelyn Ruppert. Big data: Issues for an international political sociology of data practices. *International Political Sociology*, 10(3):275–296, 2016.
- Mastercard. Mastercard API Status. <https://developer.mastercard.com/api-status>, Accessed 2024.
- Microsoft Fabric. <https://learn.microsoft.com/en-us/fabric/get-started/fabric-trial>, Accessed 2024. Microsoft Fabric documentation.

- Microsoft Power Automate. Microsoft power automate. <https://www.microsoft.com/en-gb/power-platform/products/power-automate>, Accessed 2024.
- Microsoft Power Query. Microsoft power query. <https://powerquery.microsoft.com/en-us/>, Accessed 2024.
- Patrick Montjourides. Education data in conflict-affected countries: The fifth failure? *Prospects*, 43(1):85–105, 2013.
- Robin Nixon. *Learning PHP, MySQL, JavaScript, and CSS: A step-by-step guide to creating dynamic websites*. ” O’Reilly Media, Inc.”, 2012.
- Our World in Data. OWID Catalog. <https://pypi.org/project/owid-catalog/>, Accessed 2024.
- Marco Piccioni, Carlo A. Furia, and Bertrand Meyer. An empirical study of API usability. In *2013 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 5–14. IEEE, 2013.
- Power BI. Power bi connectors. <https://learn.microsoft.com/en-us/power-query/connectors/>, Accessed 2024a.
- Power BI. Power bi service vs desktop. <https://learn.microsoft.com/en-us/power-bi/fundamentals/service-service-vs-desktop>, Accessed 2024b.
- Power BI. Power bi service gateway. <https://learn.microsoft.com/en-us/power-bi/connect-data/service-gateway-onprem>, Accessed 2024c.
- Power BI. End-user reading view. <https://learn.microsoft.com/en-us/power-bi/consumer/end-user-reading-view>, Accessed 2024d.
- PowerBI Microsoft Learn. <https://learn.microsoft.com/en-us/rest/api/power-bi/>, Accessed 2024. Power BI REST API documentation.
- Atyanti Prabaswari, Chancard Basumerda, and Bagus Utomo. The mental workload analysis of staff in study program of private educational organization. *IOP Conference Series: Materials Science and Engineering*, 528:012018, 06 2019. doi: 10.1088/1757-899X/528/1/012018.
- PyCharm Code Coverage. <https://www.jetbrains.com/help/pycharm/code-coverage.html>, Accessed 2024. PyCharm documentation.
- Qlik. Qlik sense data connectivity. https://help.qlik.com/en-US/sense/February2024/Subsystems/Hub/Content/Sense_Hub/LoadData/connect-data-sources.htm, Accessed 2024.
- REACH. Whole of afghanistan assessment 2022 key findings presentation - inter-cluster coordination team, kabul, 20 september 202. Technical report, ReliefWeb, 2022.
- Martin P. Robillard and Robert DeLine. A field study of API learning obstacles. *Empirical Software Engineering*, 2010.
- Namrata Saraogi, Diana Katharina Mayrhofer, and Husein Abdul-Hamid. Saber education management information systems country report: Afghanistan 2017. Technical

- report, World Bank Group, Washington, D.C., 2017. URL <http://documents.worldbank.org/curated/en/350071500371066946/SABER-education-management-information-systems-country-report-Afghanistan-2017>.
- Save The Children. Count Every Child. <https://www.childatlas.org/blog/count-every-child>, 2023.
- Save the Children. Majority of school girls in afghanistan missing education. <https://www.savethechildren.org.uk/news/media-centre/press-releases/majority-of-school-girls-in-afghanistan-missing-education>, 2024.
- Souhaila Serbout and Cesare Pautasso. "An Empirical Study of Web API Versioning Practices". In Irene Garrigós, Juan Manuel Murillo Rodríguez, and Manuel Wimmer, editors, *Web Engineering*, pages 303–318, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-34444-2.
- Aisha Shalash, Niveen M E Abu-Rmeileh, Dervla Kelly, and Khalifa Elmusharaf. The need for standardised methods of data collection, sharing of data and agency coordination in humanitarian settings. *BMJ Global Health*, 7(Suppl 8):e007249, 2022.
- De S Sirisuriya et al. A comparative study on web scraping. 2015.
- S M Sohan, Frank Maurer, Craig Anslow, and Martin P. Robillard. A study of the effectiveness of usage examples in REST API documentation. In *2017 IEEE symposium on visual languages and human-centric computing (VL/HCC)*, pages 53–61. IEEE, 2017.
- Siddharth Subramanian, Laura Inozemtseva, and Reid Holmes. Live API documentation. In *Proceedings of the 36th international conference on software engineering*, pages 643–652, 2014.
- Tableau. Tableau data management. <https://www.tableau.com/en-gb/products/add-ons/data-management>, Accessed 2024.
- UN Displacement Tracking Matrix. Un displacement tracking matrix. <https://dtm.iom.int/node/26946>, Accessed 2024.
- UN-OCHA Humanitarian Data Exchange Project. Un-ocha humanitarian data exchange project github. <https://github.com/OCHA-DAP>, Accessed 2024.
- UNESCO, 2021. The right to education: What’s at stake in afghanistan? - a 20-year review.
- UNESCO Institute for Statistics. The data revolution in education - paper no. 39, 2017.
- UNICEF. SDMX API Documentation. <https://data.unicef.org/sdmx-api-documentation/>, Accessed 2024a.
- UNICEF. UNICEF Afghanistan Data. <https://data.unicef.org/country/afg/>, Accessed 2024b.
- UNICEF Multiple Indicator Cluster Surveys (MICS). UNICEF MICS. <https://uis>.

- unesco.org/sites/default/files/documents/education-statistics-faq-en.pdf, 2015.
- United Nations. Goal 4. Department of Economic and Social Affairs, 2023. <https://sdgs.un.org/goals/goal4>.
- United Nations High Commissioner for Refugees (UNHCR). UNHCR Data Portal. <https://data.unhcr.org/en/about/>, Accessed 2024.
- World Bank. About the Indicators API Documentation. <https://datahelpdesk.worldbank.org/knowledgebase/articles/889392-about-the-indicators-api-documentation>, Accessed 2024a.
- World Bank. World bank open data. <https://data.worldbank.org/>, Accessed 2024b.
- World Bank Data Bank. World development indicators. <https://databank.worldbank.org/source/world-development-indicators>, Accessed 2024.
- Bo Zhao. Web scraping. *Encyclopedia of big data*, 1, 2017.

Appendix A

Participants' information sheet

This study was certified according to the Informatics Research Ethics Process, reference number 8095/773411. Please take time to read the following information carefully. You should keep this page for your records.

Who are the researchers?

Ojaswee Bajracharya, Fiona McNeill and Amanda Meyer.

What is the purpose of the study?

The purpose of the study is to assess the functionality of the live-linking system over a longer time period and gather insights and feedback regarding the system. The live linking system will help improve completeness and timeliness of getting data onto a system to be able to view data efficiently.

Why have I been asked to take part?

The aim of the study is to evaluate how effective the live-linking system is. Your contribution will help judge the efficiency of the new system across a longer time period. You also provide a different perspective, allowing for a comprehensive assessment and help improve the system.

Do I have to take part?

No – participation in this study is entirely up to you. You can withdraw from the study at any time, up until we've done the analysis without giving a reason. After this point, personal data will be deleted and anonymised data will be combined such that it is impossible to remove individual information from the analysis. Your rights will not be affected. If you wish to withdraw, contact the PI. We will keep copies of your original consent, and of your withdrawal request.

What will happen if I decide to take part?

We will ask you to create reports in the Power BI web application and ask you to check these reports every weekday for two weeks, for any updates, and to note down any changes. We will carry out a mid-study semi-structured interview at the end of the first

week. Then at the end, we will carry out a final semi-structured interview and ask you to fill out a questionnaire about the task.

Are there any risks associated with taking part?

There are no significant risks associated with participation.

Are there any benefits associated with taking part?

No.

What will happen to the results of this study?

The results of this study may be summarised in published articles, reports and presentations. Quotes or key findings will be anonymized: We will remove any information that could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research. Your data may be archived for a maximum of 4 years. All potentially identifiable data will be deleted within this timeframe if it has not already been deleted as part of anonymization.

Data protection and confidentiality.

Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a unique participant number rather than by name. Your data will only be viewed by the researcher/research team Ojaswee Bajracharya, Fiona McNeill and Amanda Meyer.

All electronic data will be stored on a password-protected encrypted computer, on the School of Informatics' secure file servers, or on the University's secure encrypted cloud storage services (DataShare, ownCloud, or Sharepoint) and all paper records will be stored in a locked filing cabinet in the PI's office. Your consent information will be kept separately from your responses in order to minimise risk.

What are my data protection rights?

The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Who can I contact?

If you have any further questions about the study, please contact the lead researcher, Ojaswee Bajracharya (s2015068@ed.ac.uk). If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint. Updated information.

If the research project changes in any way, an updated Participant Information Sheet will be made available on <http://web.inf.ed.ac.uk/infweb/research/study-updates>.

Alternative formats.

To request this document in an alternative format, such as large print or on coloured paper, please contact Ojaswee Bajracharya (s2015068@ed.ac.uk).

General information.

For general information about how we use your data, go to: edin.ac/privacy-research

Appendix B

Participants' consent form

Participant number: _____

Participant Consent Form

Project title:	Supporting humanitarian efforts in Afghanistan: Developing the ability to live-link different data sources to Power BI
Principal investigator (PI):	Ojaswee Bajracharya
Researcher:	Ojaswee Bajracharya
PI contact details:	s2015068@ed.ac.uk

By participating in the study, you agree that:

- I have read and understood the Participant Information Sheet for the above study, that I have had the opportunity to ask questions, and that any questions I had were answered to my satisfaction.
- My participation is voluntary, and that I can withdraw at any time without giving a reason. Withdrawing will not affect any of my rights.
- I consent to my anonymised data being used in academic publications and presentations.
- I understand that my anonymised data will be stored for the duration outlined in the Participant Information Sheet.

Please tick yes or no for each of these statements.

1. I allow my data to be used in future ethically approved research.

<input type="checkbox"/>	<input type="checkbox"/>
Yes	No

2. I agree to take part in this study.

<input type="checkbox"/>	<input type="checkbox"/>
Yes	No

Name of person giving consent	Date dd/mm/yy	Signature
_____	_____	_____
Name of person taking consent	Date dd/mm/yy	Signature
_____	_____	_____

Figure B.1: Participant Consent Form

Appendix C

NASA TLX

C.1 NASA TLX Scale

C.2 NASA TLX Workload Comparison Cards

Figure 8.6**NASA Task Load Index**

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date
------	------	------

Mental Demand How mentally demanding was the task?

Very Low
Very High

Physical Demand How physically demanding was the task?

Very Low
Very High

Temporal Demand How hurried or rushed was the pace of the task?

Very Low
Very High

Performance How successful were you in accomplishing what you were asked to do?

Perfect
Failure

Effort How hard did you have to work to accomplish your level of performance?

Very Low
Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed were you?

Very Low
Very High

Figure C.1: NASA TLX Scale

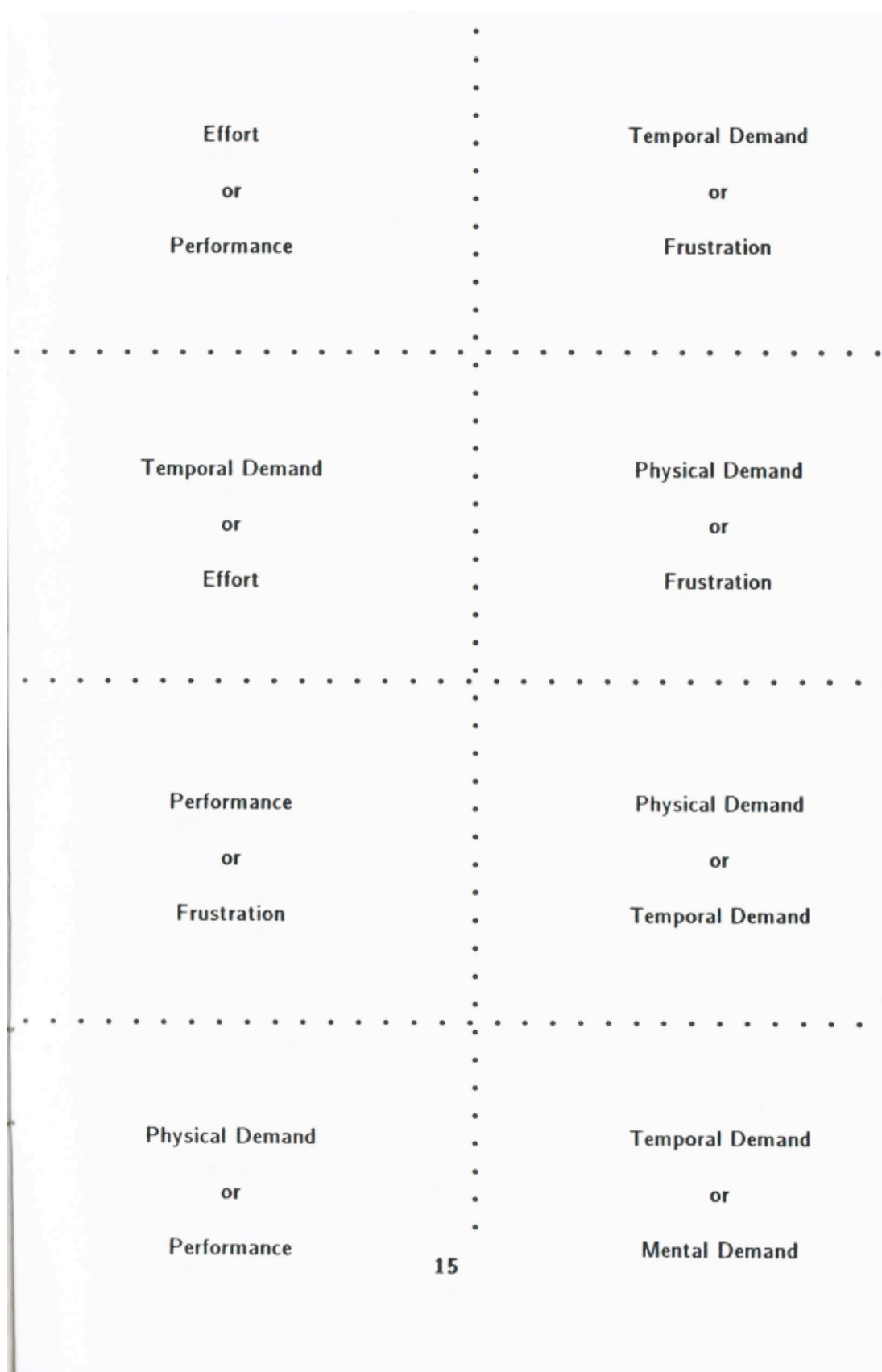


Figure C.2: NASA TLX Sources of Workload Comparison Cards Page 1

Frustration or Effort	Performance or Mental Demand
Performance or Temporal Demand	Mental Demand or Effort
Mental Demand or Physical Demand	Effort or Physical Demand
Frustration or Mental Demand	

16

Figure C.3: NASA TLX Sources of Workload Comparison Cards Page 2

Appendix D

Semi-structured Interview Questions

D.1 Mid-study Semi-structured Interview Questions

1. How has your general experience with the system been so far?
2. Have you encountered any issues?
3. What changes in the data did you notice?
4. What have you liked?
5. What do you think needs improvement?
6. Are there any additional features you think would be useful?

D.2 Post-study Semi-structured Interview Questions

1. How did you find the changes to the system?
2. Have you encountered any issues?
3. What changes in the data did you notice?
4. What have you liked?
5. What do you think needs improvement?
6. Are there any further additional features you think would be useful?