ASIA AND THE PACIFIC

Regional Innovation Centre

UNDP

Collective intelligence in action

Using machine data and insights to improve UNDP Sensemaking

Report on the Portfolio Analytics for Strategic Insights (PASI) project.

October 2022

Table of Contents

Introduction

Purpose of this report

This report was written by Pedro Parraguez Ruiz, CEO of Dataverz, a data company, with the support of Shumin Liu from the UNDP. It has been edited by the Regional Innovation Centre team and any mistakes in the editing we apologize for an can be attributed to us. The purpose of the report is to share the background to the Portfolio Analytics for Strategic Insights (PASI) project and provide a detailed report on each of the steps taken in the project. The report is technical in nature and not meant for a general audience. There is a presentation and blog that accompanies this report which is written for a more general audience.

Background of the project

To accelerate learning and impact of UNDP's work, UNDP Regional Innovation Centre (RIC) in Bangkok has been supporting Country Offices (COs) in the region of Asia and the Pacific to test out the methodology of Sensemaking.

Sensemaking helps COs establish the relevance of their portfolio of projects (do these projects make sense together? Do they take us towards our north star? Are they impactful?), Enhance coherence of their work (do these projects support each other? Where are the gaps?), bring teams together (do we know what each other is doing?) and other strategic objectives (efficiency and effectiveness of the office, offer to donors etc.) which are outlined in various blogs[1].

Sensemaking as a process is currently a well-facilitated and interactive process in the format of a two, three- or four-day workshop to reflect on the purpose of the COs programme, its relevance, its connection with the future etc. The process is designed to make the CO programming more effective, efficient, and impactful, Sensemaking is a qualitative process where people share their insights and experiences.

However, we also know that in the UNDP and Country offices there is a volume of structured and unstructured data laying in the project documents and administrative reports. The UNDP has very good data around Country Office management for managing funding, contracts, HR etc. but we wanted to know could this data be extracted in a useful way that could add value to strategy and Sensemaking. Does the quantitative data we have back up the qualitative understanding within offices? Could we build collective intelligence by exploring the hidden connections and patterns between projects by key dimensions, identifying gaps and opportunities, and providing useful insights for enhancing coherence and informing project design.

As a pilot to test this the project looks into the structured and semi-structured data from https://open.undp.org/ as well as unstructured data from Open UNDP, project documents and annual progress reports of selected projects in the UNDP Philippines, to, among other things, explore the patterns and connections between projects based on target areas of interest. This work aims to extract useful insights for the CO colleagues to better understand where their portfolio is working and identify entry points for breaking silos between teams and spurring collaboration. This work is designed to help improve Sensemaking, support better strategy and improve management decisions.

Specific Objectives of the project

- Adopt a data-driven approach to provide useful intelligence for Sensemaking that aims at enhancing portfolio coherence and accelerate institutional learning

[1] https://undp-ric.medium.com/

- Explore and identify the hidden connections and patterns among projects with the semi-structured and unstructured data in the organization.
- Enhance the UNDP Country Office's knowledge and capacity in gaining basic understanding of applying Artificial Intelligence for development work, including how it works, what it can and cannot do, potential bias and limitation and implication of ethical use of AI

Scope

As the first pilot to test out the portfolio analysis in UNDP, this project works on the project documents and administrative reports from the Philippines doing this in close collaboration with RIC and UNDP Philippines. In addition, and when relevant, we pulled global UNDP project data to test scalability potential and to run comparative analyses.

Research questions

The following questions emerged from workshops and meetings between key project stakeholders, primarily UNDP colleagues in the Philippines country office and the UNDP Regional Innovation Centre (RIC) in Bangkok.

1. **Development challenges:** What does CO portfolio look like in terms of the development challenges projects working on by geographic location, partners, investment, and interventions? How does the gender lens is being embedded? Is there a COVID-19 portfolio?
2. **How**: What are the opportunities for project teams to collaborate with and learn from each other? What are the entry points based on thematic areas, approaches, and capacity?
3. **Partners and stakeholders:** How do the mapping of stakeholders and partners look like by different types of relationship (such as donors, implementation partners, beneficiaries, other stakeholders)?
4. **Effect**: What results/impact are the projects aiming to achieve? How do they contribute to the CPD outputs? How does gender equality be reflected in the outcome/outcome measurement?
5. **COVID-19**: How does COVID-19 impact our portfolio? What are the changes we made to cope with the COVID-19 crisis? What are the lessons learned?

It is worth noting that the questions are not independent from each other, rather each of them poses an emphasis (or lens) through which to explore and make sense of the project portfolio, hence their answers are sometimes intertwined.

To simplify the understanding of each of the exploration points that underlie the research questions, an alternative way of describing the research angles taken is through the following analytical lenses:

- **Topic-discovery and classification lens:** insights about the underlying network of themes/topics that serve to connect and differentiate projects within the portfolio. *This lens is strongly connected to research questions #1 and #2.*
- **COVID-19 lens** insights about the connection between projects responses to COVID-19 (or lack of response) and the characteristics of those projects within the portfolio. *This lens is strongly connected to research question #5*
- **Gender lens**: insights about the connection between projects acknowledgements/references to gender-related issues (or lack of such references) and the characteristics of those projects within the portfolio. *This lens is strongly connected to research question #1 and it is relevant in the context of RQ #4.*
- **Geographical lens**: insights about the connection between the geographical footprint of projects and the characteristics of those projects within the portfolio. *This lens is strongly connected to research questions #1 and #2.*

- **Partners and stakeholders' lens**: insights about the connections between the type of partners and stakeholders involved in projects and the characteristics of those projects within the portfolio. *This lens is strongly connected to research question #3.*

What we didn't do

We worked with the Philippines Country Office and presented the work, these inputs were used during a Sensemaking process in the Philippines Country Office in 2021, however we didn't map or track if better or different decisions were made because of this work, we also didn't work out a way to mainstream this into Sensemaking. This was due to changes in personal in the Philippines Country Office and at the Regional Innovation Centre but also funding for this work. Ultimately this project was a start. We showed how the data could be gathered and organized in a useful way, we uncovered some unusual patterns that humans alone couldn't see but we didn't show how this could be mainstreamed into decision making or into Sensemaking.

High level project findings related to the Philippines Country Office

The purpose of this project was to extract insights from the data available both structured and unstructured to see what emerged. Then this data would be married with the qualitative findings from the people who are part of Sensemaking. Here is what the project discovered about the Philippines Country Office.

Themes and topics

- "Planet" and "Peace" represent close to 50% of the overall project portfolio of the office
- The class "Others" comes in third place, and within this class, "Cross-cutters and enablers" form a significant part of the offering, revealing the relative importance of this kind of general capacity generation projects.
- "Health" related categories seem to be one of the most challenging categories to classify. Classification issues relate to the use of terms such as "health", "well-being" or "recovery" outside of the healthcare context. This can be evidenced for example with topics such as the economy or natural disasters where terms like healthy economy or recovery are common. This leads to some false positives associated with the "health" category. Despite a limited number of misclassified nodes, the nodes` position in the graph remains a good indication of similarity. This is because we use multiple language models to calculate the position of the nodes and identify clusters, not just the core service offering classification.

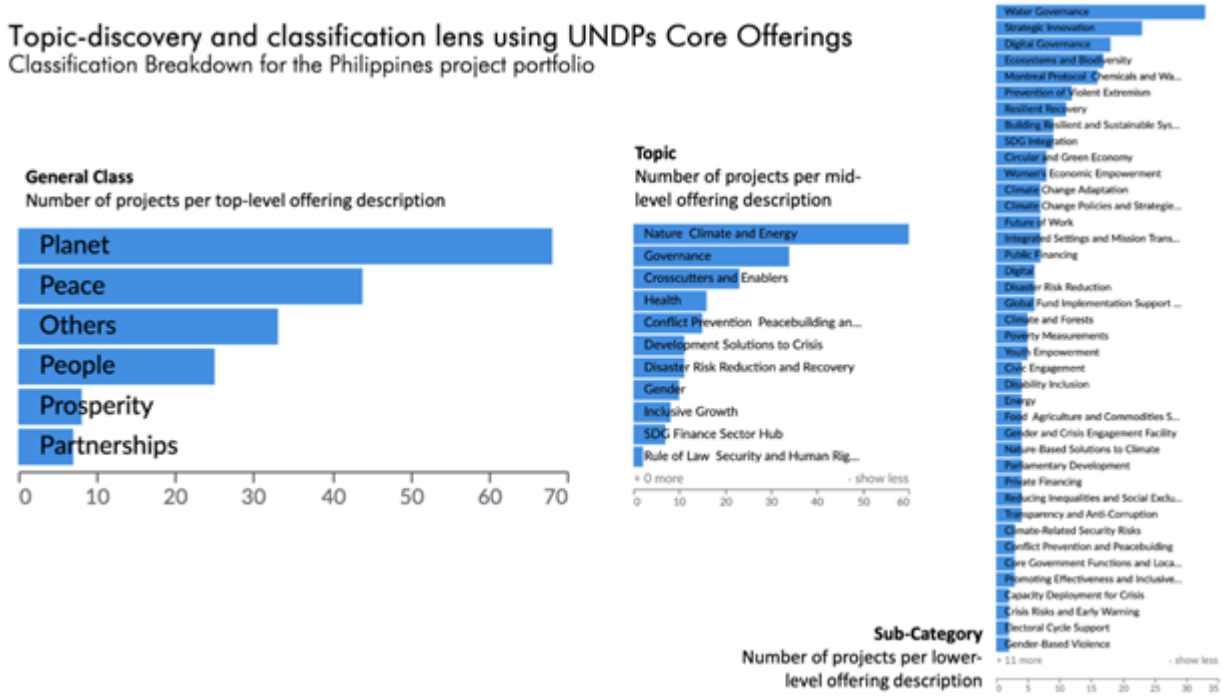

Figure 1: UNDPs Core Offerings Classification Breakdown for the Philippines project portfolio

- Despite their relatively smaller size within the portfolio, the offerings termed "Prosperity" and "Partnerships" play a central brokerage role in the themes that they cover. Projects tagged under these offerings appear to be well suited to connect different clusters, such as creating

connections between projects under the offerings of “Planet” and “Cross cutters and enablers” (which is the main topic within “Others”)

- The boundary between the offerings of people and peace is the one that shows the most instances of project similarities, meaning that thematically these two clusters tend to coincide more (and therefore it might be easier to set up collaborations within shared areas of work). The main bridges between these two clusters are the topics of “Governance” and “Gender”
- The cluster that appears to be thematically most disconnected from the rest is the one with offerings labelled “Planet”. The most direct bridges between the “Planet” cluster and the rest of the network are via “Partnerships”, “Prosperity” and via” Cross cutters and enablers” (under the main offering labelled “Others”).

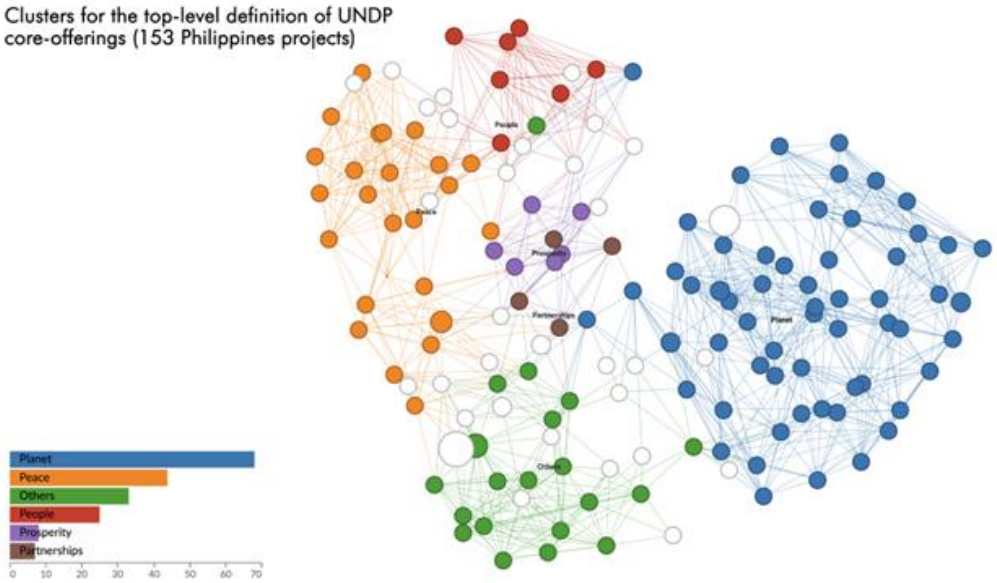

Figure 2:Clusters of projects based on semantic similarities and their relationship with the core offerings.

Identification of potential bridges/brokers between different UNDP offerings

- The network graph shows the subset of projects that have active years from 2019 onwards (50 projects in total). The objective is focusing on identifying projects that might be particularly well positioned to act as bridges between different areas. The quickest way of identifying those projects in the graph is by narrowing down on projects that are colored white (meaning that they have been labelled under more than one top level core offering) as well as targeting projects that in the graph appear to be at the intersection of different clusters (right in between groups representing different core offerings).
- Using the heuristics described above, a project like “Philippines: Low Emission Capacity Building Project” appears to have some characteristics that, in relative terms within the portfolio, provide it with a good position to act as a bridge. Indeed, a closer inspection reveals that this project seems to be a good candidate to act as a knowledge bridge within the portfolio given some of its self-reported characteristics, namely:
    - Reported SDGs: Sustainable cities and communities + No Poverty SDGs that are relevant in multiple challenge areas.
    - Reported “How’s”: Policy Advice, Normative Support, and Institutional Mechanism and System Building - “How’s” that are inherently about knowledge sharing and integrative system solutions.

- The project includes a wide range of partners from the private sector, national government, sub-national government, and NGOs.
- In “Who” the project reports addressing a wide variety of beneficiaries (6 in total)
- The project reports a focus on partnerships, economic development, and climate change all of which have a wide relevance.

Figure 3: Illustrating connections between core offerings.

Philippines UNDP core offerings and their regional context

- For the overall portfolio (all years, all projects), some of the most salient and robust differences between the portfolio of the Philippines and the rest of Regional Bureau for Asia and the Pacific (RBAP) are:
  - In terms of the top-level classification, the Philippines project portfolio has a larger proportion of projects under the “Planet” offering and a lower proportion of projects under the “Peace” and “Partnerships” offering.
  - At lower classification levels, the largest difference is in” Transparency and Anti-Corruption” where RBAP has almost 9% of the projects tagged under this offering and the Philippines has slightly under 3% of the projects tagged under this offering.
  - Other statistically significant differences between the portfolios exist in the offerings “Digital Governance”, “Private Financing” and “Digital”. In all of which the proportion of projects tagged under those offerings is higher in the rest of RBAP than in the Philippines.
- The differences are reduced or reversed when we examine projects after 2016, when the proportion of “Planet” projects in the Philippines ‘portfolio goes down relative to the rest.

Geographical context for the core offerings

- When looking at the geographical distribution of the labelled offerings in Philippines` project portfolio, we see evidence of clear geographical specialization.
- The south of the country appears largely connected to places where most of the projects are focused on “Peace”

- The west part of the center appears largely connected to places where most of the projects are focused on" Planet" related offerings.
- The east part of the center is divided in two, a lower and an upper side.
    - The upper side is largely connected to places where most of the projects are focused on "Others" (largely crosscutters and enablers)
    - The lower side has more of a mix, combining "People" and "Planet"
- Places in the north tend to have a more diverse mix of project offerings (represented in white in the map). In the north the projects tend to combine a mix between "Peace" and "People" alongside "Prosperity" and "Cross cutters and enablers".

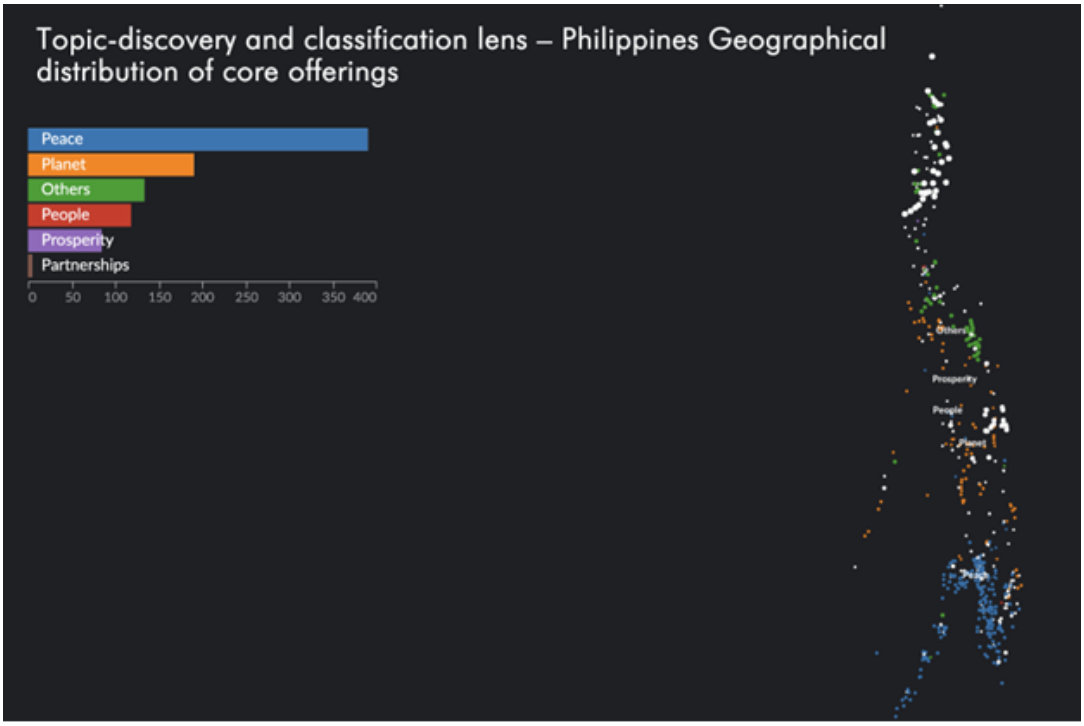

Figure 4: Illustrates regional clusters of projects

Trends and evolution of the core offerings

- The most salient trend in terms of the proportion of offerings over time (based on number of projects per active year) is the relative decline of projects labelled under the "Planet" offering and the marked increase of projects labelled under "Others", category that includes mostly "Cross-cutters and enablers" as well as "Development Solutions to Crisis".
- From 2018 onwards we see evidence of an overall increase in the diversification of the offerings in the portfolio of active projects. Instead of having a large proportion of projects in "Planet" (that at points reached almost 50%) the overall distribution of the offerings becomes more diverse. This is true even as the category "Others" (that is intrinsically more heterogeneous) grows larger than "Planet".
- At a more granular level, the growth in offerings labelled as "Others" appears to be driven by offerings such as "Strategic Innovation", "SDG Integration", and "Digital".

Results of text mining PDFs for the Philippines

- To test the usefulness and feasibility of using text extracted directly from project-related PDF documents we took 44 recent Philippine projects and mined a selection of their PDF documents.
- The data extracted from the PDFs was later combined with the data from Open UNDP to further enrich the pool of information available per project
- Two areas of immediate interest for portfolio insights that are only available in the PDF documents are reported partners (the names and types of the organizations) and reported innovations. To explore these two additional angles, we created a dashboard focused on the 44 projects with data from the PDFs to explore them.
- In reported innovations, it is noticeable the relatively high share of "Mobile-based feedback mechanisms" within this portfolio subset. Other innovations frequently reported include "New and Emergent Data", "Human-Centered Design", and "Real Time Monitoring
- In terms of reported partners, the main challenge is name disambiguation/reconciliation. We manage to merge with computational methods some of the most common name variations, but the use of ambiguous acronyms mixed with full and partial names means, that now, to fully merge and consolidate names a larger effort is needed that combines both the help of project members and additional computational methods.

COVID-19 lens

Insights about the connection between projects responses to COVID-19 (or lack of response) and the characteristics of those projects within the portfolio.

Marker inputs used to tag the projects:

- Output Marker" COVID-19 Response" (453 global projects)
- Global text mining using terms like "Corona" and "COVID" within open and closed descriptions (Open UNDP) incl. markers and results

Distribution of COVID response within the project portfolio

- To evaluate the distribution of COVID-19 related actions or references within the project portfolio we overlay the COVID marker on top of the Philippines project network, selecting only projects with active years that include 2020 onwards.
- We found COVID markers on almost 50% of the projects. This means that close to half of the projects responded to COVID by adapting or including outputs that seek to address COVID either to adapt the project to the new pandemic situation or to offer beneficiaries COVID-related solutions.
- Within this limited subset of projects there is some evidence that certain types of projects were more responsive than others. This is shown in this graph in the large orange cluster demarcated with red. We will investigate the specific characteristics later this report.
- Using the network graph, we can also see projects that appear clearly outside the main cluster. A good example is the project "Strengthening the Marine Protected Area System to Conservation". The type of projects that show a response to COVID despite being outside the main cluster are interesting because they could show strategies and/or share the knowledge that allow them to respond despite their characteristics. In this example, this project could collaborate with semantically similar projects within the "Planet" area that for the most part show a low response to COVID.

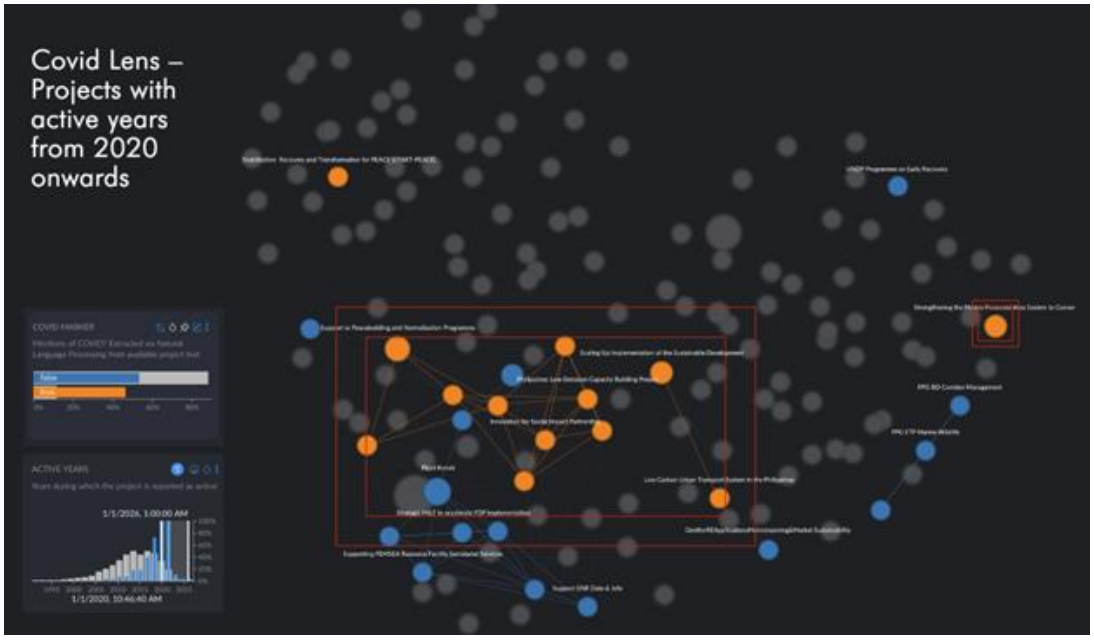

Figure 5: Illustration of covid response projects

Comparisons between projects with and without COVID-19 marker

- Besides the positions within the network graph, we can use other markers to understand if in general projects that showed a response in connection to COVID seem to be alike or not. As expected, we observe that there are some attributes that seem to increase the chances that a project responds to COVID.
- Some of the project attributes that have a positive correlation with positive COVID response (COVID marker = "true") are:
    - Being labelled under the top-level UNDP offerings of "Peace", "People" or "Partnerships". Being level under the mid-level UNDP offerings of "Governance", "Health", or "Development Solutions to Crisis". Being labelled under the lower–level UNDP offerings of "Digital Governance", "Circular and Green Economy", "Capacity Deployment for Crisis", "Global Fund Implementation Support and Health Procurement", or "Integrated Settings and Mission Transitions".
    - Projects with larger budgets show more responsiveness.
    - "How's" such as "Capacity Development / Technical Assistance", "Institutional Mechanism and System Building", "Data Collection and Analysis" and "Convening / Partnerships / Knowledge Sharing" show more responsiveness.
    - Focus areas such as "Accelerate structural transformations", "Eradicate poverty in all its forms and dimensions" and "Build resilience to shocks and crisis" show more responsiveness.

Geographical comparisons between projects with and without COVID-19 marker

- We observe that certain geographical clusters seem to be more likely to show a response to COVID-19 than others.
- Overall, the south and the west part of the center show a higher response than the north and the east side of the center.

- These differences are probably connected to the already existent geographical specialization of the projects that was previously discussed which we have found in turn that it is also correlated to COVID-19 response.

Regional context (Regional Bureau for Asia and the Pacific – RBAP)

- When compared to the rest of RBAP, the Philippines shows a higher proportion of geographical places where we found projects with COVID markers.
- Although in terms of number of projects the rest of RBAP has more than 10 times the number of projects compared to the Philippines, the geographical footprint of projects within the Philippines is the largest in terms of number of registered places.] As an example, in Open UNDP the Philippines has 584 registered places and 153 projects. In turn China has 172 registered places and 217 projects. This might be the result of registering at different levels of granularity between the different country offices (COs) or a real difference in terms of geographical footprint. The interpretation of any comparison between COs will therefore depend on the accuracy and completeness of these geographical markers.

Gender Lens

Insights about the connection between projects´ acknowledgements/references to gender-related issues (or lack of such references) and the characteristics of those projects within the portfolio.
Marker inputs:

- Declared Gender "Policy Significance"
- Text mining using terms like "gender" and "female", "woman", etc. within open and closed descriptions (Open UNDP) incl. other markers and results

Distribution of Gender within the project portfolio

- Unlike the results connected to the COVID marker, the Gender marker does not have the same clear clustering or distribution in the network graph of semantic similarities. However, a closer inspection does reveal that the cluster that contains mostly "Planet" related projects has in relative terms fewer projects referring directly to gender. In turn, the clusters that contain offerings labelled as "Others", "People" and "Prosperity" have proportionally more projects referring directly to gender than what would be expected if the distribution was homogeneous.
- At a more granular level of the core UNDP offerings, projects under the following labels have a higher proportion of the gender marker than average:
    - "Development Solutions to Crisis", "Inclusive Growth", "Health", "Gender" (unsurprisingly), "Integrated Settings and Mission Transitions", "Future of Work", "Promoting Effectiveness and Inclusive Governance for HIV and Health".
- In turn, projects under some of the following labels have a lower proportion of the gender marker than average:
    - "Disaster Risk Reduction and Recovery", "Nature Climate and Energy", "Climate Change Adaptation"

Comparisons between projects with and without gender marker

- Some of the declared "How's" that most commonly have references to gender are "Institutional Mechanism and System Building", "Direct support / Service Delivery", "Policy Advice", and "Data Collection and Analysis".
- Partners categorized as "private sector", "subnational government", "research institution" and "donor government" show a higher proportion of gender markers.
- Projects with a larger budget (above 1M USD) shows a higher proportion of projects with gender markers.

- The signature solution" Keeping people out of poverty" is the one with the highest proportion of projects with gender markers.
- Projects with references to COVID are also more likely to make references to gender.
- The reported SDG also has some correlation to the gender marker (i.e., work on some SDGs is more likely to be accompanied by a positive gender marker and work on other SDGs is more likely to be accompanied by a negative gender marker).

Geographical comparisons between projects with and without gender marker

- There is evidence of a relatively small effect between geography and the gender marker.
- The northern part of the country has proportionally fewer places with projects that have a negative gender marker than the south.
- Overall, we see broad geographical coverage of places with projects that have a positive gender marker.

Regional context (Regional Bureau for Asia and the Pacific – RBAP)

- Countries like India, Pakistan and Indonesia show a higher proportion of places with projects that have positive gender markers (compared with the Philippines).
- China shows a lower proportion of places with projects that have negative gender markers (compared with the Philippines).
- Overall, the Philippines ranks somewhere in the middle when it comes to geographical coverage of the gender marker.

Regional comparisons including trends

- We can observe a consistent and sustained increase in the number of projects with positive gender markers.
- Projects with an active year in 2019 reached a 50% of positive gender markers. After that year, the proportion of projects with positive gender markers is larger than the proportion of projects with negative gender markers.
- This trend is not exclusive to the Philippines and can be seen in most countries. Compared to the strength of the trend in most countries the Philippines follows the average tendency.

Partners and stakeholder's lens

- Older projects have a larger proportion of national and subnational governments as partners. In turn more recent projects see an increase in the participation of private sector and academic institutions.
- There is some degree of specialization between partners and the variable "Who" (the beneficiaries). For example, subnational government partners are, in relative terms, top of the list in projects for "Persons negatively affected by armed conflict or violence" and bottom of the list in "People living in urban areas".
- Likewise, there is some degree of specialization between partners and the variable "SDG". NGOs/NCOs are top of the list in projects connected to "Peace justice and strong institutions" and bottom of the list on projects connected to partnerships for the goals.

Caveats on the findings

- The results are highly dependent on the accuracy and completeness of the data provided by the projects in the form of data within Open UNDP and in some cases their project-related documents.
- The heterogeneity in the accuracy and completeness of the data across projects means that a qualitative validation step is always required before drawing direct conclusions from the data or the analytics.
- Machine understanding of human language is a challenging and constantly evolving field. The following results and interpretations leverage advanced methods with the purpose of showing emergent opportunities to scale up this kind of support and it has been deployed as a pilot to be further evaluated and improved over time.
- We have mapped specific research questions to generic questions and then the generic questions to lenses (see figure below).

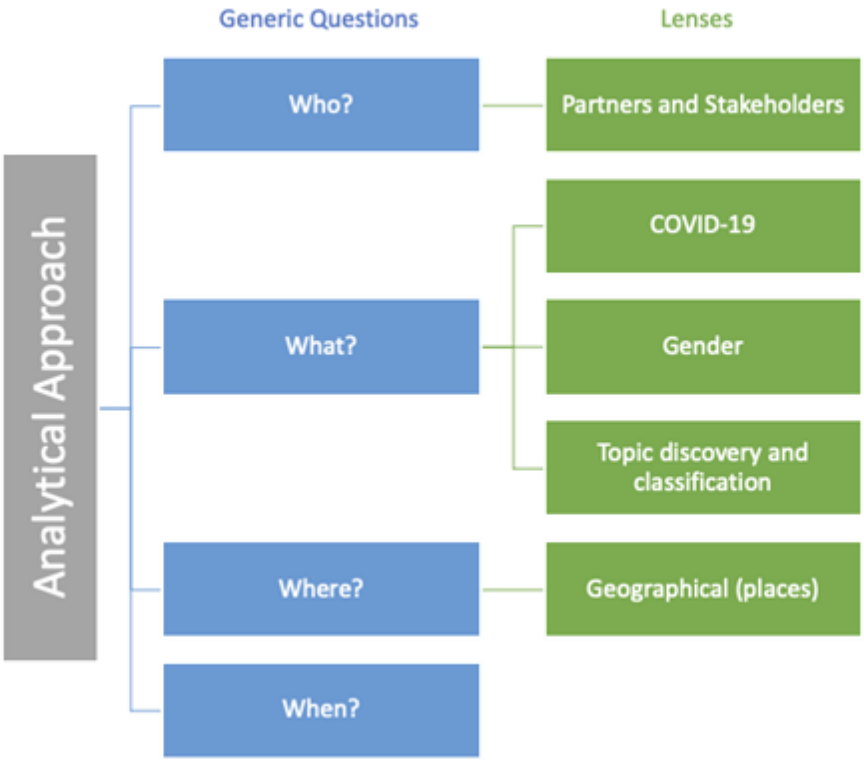

Figure 6: Connecting research questions to lenses.

Improvement recommendations

Throughout the process we learnt a lot about the quality and availability of data at the UNDP and how easy it is to use for Sensemaking and for other purposes. Below are some key takeaways for the UNDP data and strategy/innovation teams.

Strengths and weaknesses of the source data

- Strengths
  - Most of the key metadata for all projects is available via standardized API at https://open.undp.org/ under an open Creative Commons Attribution 3.0 IGO License.
  - The data obtained through the API is well organised and normalised. Having an open API also helps to set up a process that can make updates significantly easier to integrate within the data pipeline.
- Weaknesses
  - The project documents listed in Open UNDP are often mislabeled, the PDF files vary greatly in terms of quality, and all the key project documents are not always uploaded.
  - The number of pages of unstructured text within the PDF documents and the variability between the documents makes it difficult to scale up their usage.
  - Some information, like partners, innovations and gender issues are only reported within the PDFs. This otherwise structured data would be very useful if it was made available via Open UNDP´s API.
  - Data on partners (the names of the organisations that act as local project partners) are not normalised and no ID or pseudo-id (like a website) is provided.
  - The API documentation is not always sufficient to appropriately interpret the fields.
  - Geographical locations connected to the document are reported for most of the projects, however, it seems that the level of detail of the coordinates and the completeness of the reported geographical places seems to change between projects and country offices.
  - The definition of project budget (as well as resources) varies within the UNDP. The API documentation does not specify what is the budget definition that is being used for the reported values.

Limitations and other notes related to the applied methods

- To generate the network graph and define the positions of the nodes (in this case the projects) we need to compress to two dimensions (X and Y) numerous dimensions (most of the time hundreds) each of which represents project features such as attributes within the project description. This means that the two-dimensional representation of the network graph won't be able to show the exact relative distance between each project in the graph. This is like the problem of representing the globe in two dimensional graphs. For this reason, it is important to understand the graph as a visual aid only and not as a perfect representation of the portfolio. Having said that, even if for individual projects the representation might not be entirely accurate when it comes to clusters or larger segments of projects the representation tends to be more robust.
- The qualitative interpretation and validation of the quantitative results should be an irreplaceable part of an augmented sensemaking process. Since the project portfolio represented on the tool has not perfect data and the computational methods are fallible users should be encouraged to approach results critically and challenge them when needed.

- To provide a way to classify projects based on UNDP´s core offerings, we developed a method (see section on computational methods) that could help us categorize projects at scale based on the official descriptions of UNDP´s core offerings. Since beyond the description provided per offering, we do not have a "ground truth" against which to compare or validate our results our labels should only be used as a reference and a starting point from which to refine the classifiers. Furthermore, to keep the pipeline aligned to traditional classification rules, we use only the top two weighted classifications of UNDP offerings (per hierarchical level and per project).

Sensemaking Applications

At a general level, having a better understanding of the project portfolio for any potential decision that might come down the line is a good thing. However, without more specific use cases the risk is that the exercise becomes too exploratory, and we miss concrete applications where augmented sensemaking can make an important difference given its increased breadth and scale.

Some of the applications that we have explored and that seem well suited for the augmented sensemaking presented here are:

- **Identification of latent collaboration potential**.
  Identify groups of projects that share characteristics that make them particularly well suited for collaboration and/or knowledge sharing within a particular cluster of similar projects.
- **Identification of projects that can act as knowledge brokers/bridges**.
  Identify the projects that have unusual combinations of characteristics that cross between different silos/clusters and that make them particularly well suited as bridges or brokers between groups of projects that seem to be significantly different.
- **Assess the organic evolution of the portfolio VS the strategic intent**
  As time passes operational tactics and decisions can distance the original strategy behind the project portfolio of a country office or region. This might be a good natural adaptation or a sign of an unwanted disconnect. The ability to scan and process not only the original project description but also the descriptions of the ongoing outputs and reported results helps to assess the direction that the portfolio has taken and if needed plan interventions.
- **Identify capability gaps or strengths**
  Projects exist within an evolving context where geography, emergent crises and previously unavailable possibilities require constantly assessing capability gaps and the strengths available on the ground. Since capabilities and strengths are not a result of only individual projects but rather of a large portfolio of them, the quantification and mapping of the portfolio at scale can help to work with new types of metrics and indicators that represent more than just the linear sum of projects 'characteristics.

Further Recommendations

- Complement PDF documentation with structured forms that can capture the reported project data within a JSON-based document database before turning the project documents into PDFs
- Data normalisation, e.g., partners including at least websites or some other form of ID system to avoid duplication.
- This project could be used as an opportunity to further embed augmented portfolio capacities within UNDP both in terms of tool usage (e.g., training super-users and facilitators) as well as in terms of tool creation/adaptation (e.g., using the knowledge graph in other applications such as the automatic suggestion of classifications/tags of projects during the submission). The most pressing item is having a larger base of expert users that can feel empowered and help other colleagues that might benefit from the type of insights that the tool can provide.

Potential next steps

- Include and extend the analysis of CO-level information (e.g., ROAR and mini-ROAR) and seek to replicate their data acquisition format/structure at the project-level and bring in new data from other sources such as capacity mapping
- Integrate within the code-base the efforts from the SDG lab https://sdgailab.org/ team, particularly the SDG labels (https://osdg.ai/) which would allow us to categorize SDGs at a more granular level and also would help us to compare the self-reported SDGs against the machine labels.
- To facilitate the communication of the tool and the adoption within UNDP more efforts need to be done to "translate" the different features, categories, and possibilities that the tool provides into a more general UNDP language.
- Compare the development challenges UNDP aspires to address with the actual impact created with existing resources, capacity, and strategies, and identify the gaps if any.
- Seek ways to include data related to actual collaborations/interactions between projects in the portfolio so that a semantic similarity VS actual collaboration gap analysis can be incorporated within the dashboards.
- Extend the usage of the more advanced graph queries directly on the graph database to power/super users.
- Include predictive analytics and other analytics that connect portfolio inputs, outputs, and outcomes.
- If more project-level data that so far only exists within the PDFs is standardized/normalized and made available via an API or data dump, include this new data in the next generation of the portfolio analytics.
- Include/connect UNDP team member profiles associated with each project to greatly enrich and extend the capabilities of the current version of the portfolio analytics.
- Start deploying the tool in real sensemaking exercises first with trained facilitators that can make the best out of the tool and document issues or limitations.
- Further iterations of this project could expand the regional scope and explore different portfolio segmentation (e.g., all projects around the world working on digital solutions or projects within a specific region or year range).

Appendices

Tasks

The following tasks were set at the beginning of the project and acted as a guideline during the development process.

- Co-design the methodology, process and workplan for the assignment with the technical working group (TWG) from RIC and UNDP Philippines. The data science team should lead the technical part, propose suitable techniques for the analysis, and draft the research framework to conduct the assignment. The TWG will articulate the objectives of the analysis, identify key areas to focus, and provide feedback.
- Review project document(s) and project progress reports and have consultations with the TWG to understand the structure and meaning of data represented in these documents in relation to UNDP Programme Policies and Operations procedures.
- Extract text and relevant data from Open UNDP, project document(s) and project progress reports (in the format of scanned PDF) against the structure of data frame defined by the TWG and the data science team.
- Conduct text mining, network analysis/topological analysis and other relevant data analysis with the support of machine learning (the models should be built for the UNDP Philippines's context) to identify and visualize connections, patterns and clusters among projects based on key dimensions such as levers of change, thematic areas, partnerships, capacities, and others (defined by the TWG). It shall be able to produce a high-level overview of the patterns as well as logical breaks of the same while also providing disaggregated details of projects when zooming in for the purpose of enhanced programme/project monitoring and oversight.
- Develop quick prototypes to test out different types of possible data analysis and test them with the TWG. The final data analysis should be captured in an interactive visual dashboard.
- Adopt an agile and iterative approach that accommodates UNDP's needs and allows the TWG to provide timely feedback on the data prototypes and the dashboard. The data science team shall be able to adopt the feedback and iterate the data analysis.
- Articulate the gist of the methodology and process for the TWG and support the TWG to interpret the analysis. Through this assignment, the data science team shall support UNDP on capacity building so that the TWG can learn the basic logic of how different data science techniques work in this case, the advantages and potential limitations/bias, implication of responsible use of data, and how to translate the results from data analysis into useful insights.
- Conduct learning sessions and webinars with UNDP RIC, UNDP Philippines, and other relevant Country Offices to explain the methodologies (in plain language), demonstrate the progress and results of the analysis and provide guidance on generating useful intelligence.
- Develop a brief report to capture the key findings and learnings (with visualization) and an associated blog post together with the TWG.

Related Initiatives

The following are some close initiatives that are related to the work developed during this project.

- **Open UNDP**: https://open.undp.org/projects.] Essential source of data and provides complementary analysis and dashboards including analytics at the country and donor-level.
- **The SDG AI Lab**: https://sdgailab.org/. AI-based SDG classification that is highly complementary and aligned to the efforts within the context of this project. Potential next steps could include integrating their machine learning classifications within the project portfolio analytics.

- Previous and current work from the UNDP Regional Innovation Centre (RIC) in Bangkok and other UNDP initiatives on portfolio sensemaking.[2]
- Internal portfolio analytics report developed within the UNDP to track and analyze projects.

Dashboards

All dashboards are accessible via https://undp.gitbook.io/pasi/prototypes-of-data-analysis

Documentation & Tutorials

Additional documentation and tutorials can be found at: https://undp.gitbook.io/pasi
Editing access to the Gitbook is available on request for UNDP users that want to contribute to the documentation.

Development Process and methodology

We managed and structured the project activities following an agile version of the system engineering V-Model for software development[3].

For this, we divided activities into three work packages (WPs):

- **WP1, System-level design:** This work package is focused on fine tuning and validating the overall system architecture, ensuring that all requirements, project information and constraints have been gathered and taken in consideration. Most activities within this package were performed during the first phase of the development process. This WP includes as the main deliverable *"Assignment proposal including methodologies and a detailed work plan with tasks and timelines"*
- **WP2, Module-level design and development:** This work package is focused on designing and developing each of the modules required for the operation of the system and it is where most of the software development work occurs. This WP includes as main deliverables "*Text mining, network analysis/topological analysis and other relevant data analysis to identify and visualize connections, patterns or break of them, clusters among projects based on key dimensions*" that is subdivided into" Prototypes of data analysis" and "Interactive dashboard of final data analysis".
- **WP3, System integration, testing, deployment, and training:** This work package is focused on integrating each of the modules of the system, testing their correct work, deploying the platform within UNDP´s premises/systems, documentation and training the users and the super users. This work package includes *"Online learning sessions and webinars to explain the methodologies, demonstrate the progress and results of the analysis, and provide guidance on generating useful intelligence"* and "*A brief report to capture the key findings and learnings (with visualization) and an associated blog post together with the TWG"*

Additionally, throughout the development process, there were periodic validation activities to contrast the design set in WP1 against the results obtained in WP2 and WP3, so that early corrective measures can be taken in case deviations arise.

[2] See sensemaking posts in https://undp-ric.medium.com/ for examples.
[3] http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.170.5689

Methodology

The figure below provides an overview of the project framework applied to interpret the project requirements and guide the overall design process. This framework is inspired by design thinking methodology and the principles of the agile software manifesto. This methodology is also compatible with sensemaking strategies as it allows for the inclusion of multiple viewpoints and stakeholders during the development process.[4]

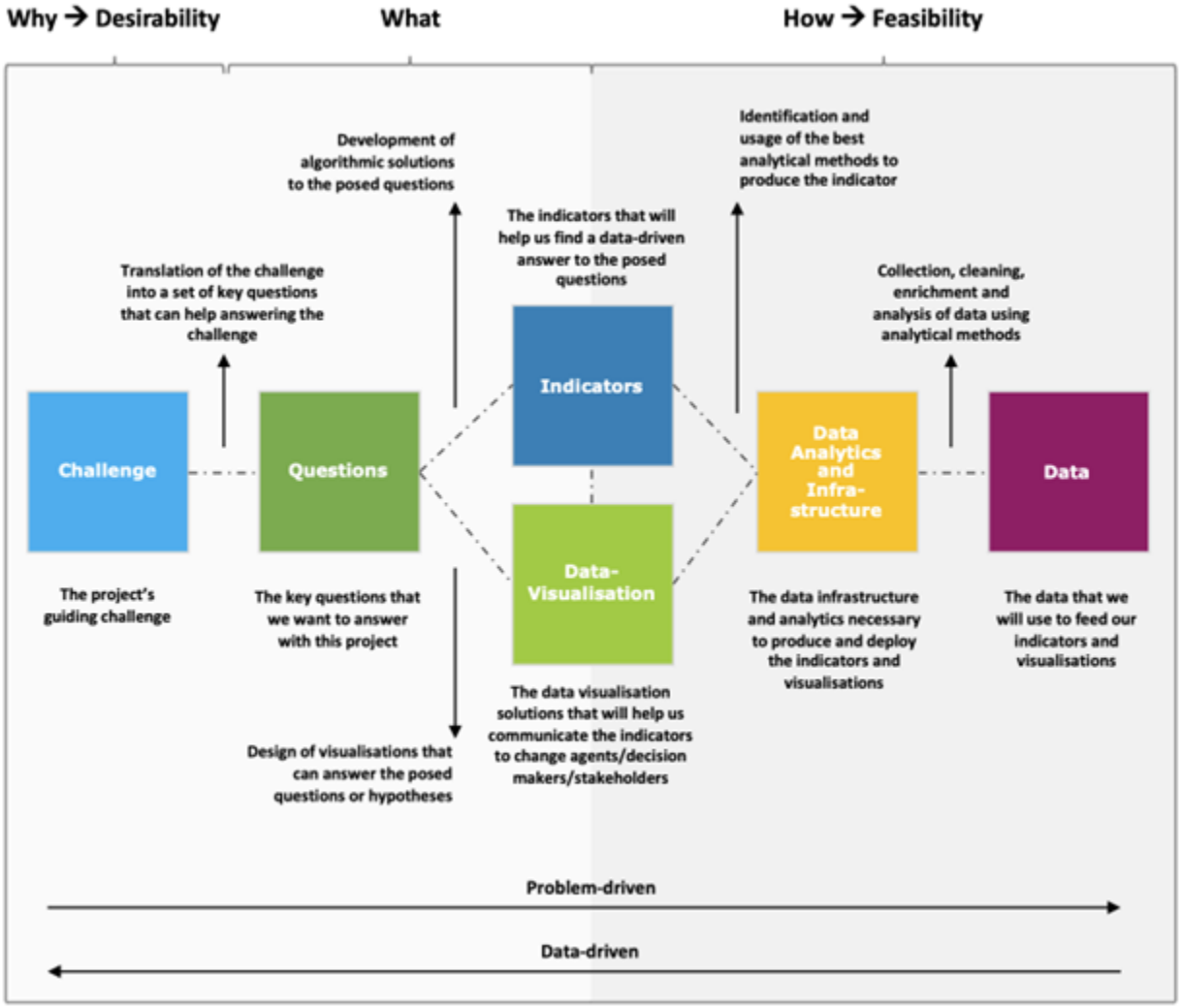

Figure 7: Overall design framework

Reflections on the application of the methodology and the iterative implementation of the development process

The process was punctuated and enriched by a series of iterations cycles that had as a main source of inputs the feedback, questions, ideas, and comments provided by a diverse set of UNDP staff that kindly volunteered their time and that also showed interest in the project. In total there were more than 24 meetings between Dataverz and the UNDP. While most of the meetings were focused one-on-one video-conference sessions. We also had two formal (online) workshops as well as learning/training sessions and presentations to a broader audience within UNDP with the objective of socializing the

[4] More details of this methodology are available here: https://medium.com/@EURITO/the-cart-before-the-horse-in-data-science-projects-back-to-basics-961c908b1796

results and getting early indications of potential use cases and areas of particular interest. The workshops were primarily focused on requirement elicitation (e.g., identifying key research questions and areas of interest) as well as having a better understanding about the potential users.

This process allowed us to move from the first rapid prototypes to more consolidated dashboards. An additional tool that facilitated the process was the open Gitbook[5] where we incrementally documented the process along the way and kept the related material so that people joining in later in the process could have a better understanding of the projects and its development.

Data

The main data source for this project is the official UNDP open data portal for projects: https://open.undp.org/projects particularly the following json endpoints:

- Individual Project Data: https://api.open.undp.org/api/projects/{project - id}.json
- Project Summaries: https://api.open.undp.org/api/project_summary_{year}.json
- Operating Unit Data: https://api.open.undp.org/api/units/{operating - unit}.json
- Operating Unit Index: https://api.open.undp.org/api/units/operating-unit-index.json
- Sublocation Location Index: https://api.open.undp.org/api/sub-location-index.json
- Region Index: https://api.open.undp.org/api/region-index.json
- Donor Index: https://api.open.undp.org/api/donor-index.json
- Donor by Country Index: https://api.open.undp.org/api/donor-country-index.json
- Focus Area Index: https://api.open.undp.org/api/focus-area-index.json
- Aid Classification Index: https://api.open.undp.org/api/crs-index.json
- SDG Index: https://api.open.undp.org/api/sdg-index.json
- Individual Output Data: https://api.open.undp.org/api/outputs/{output - id}.json
- SDG Target index: https://api.open.undp.org/api/target-index.json
- Individual SDG Target index: https://api.open.undp.org/api/target-index/{sdg - id}.json
- Signature solution index: https://api.open.undp.org/api/signature-solutions-index.json
- Our Approaches index: https://api.open.undp.org/api/our-approaches-index.json
- Project Data: https://api.open.undp.org/api/project_list/

We also make use of the internal API for the website itself: https://api.open.undp.org/api/v1/

In terms of unstructured text from project descriptions we gathered data from PDF files associated to project descriptions of interest for a total of 44 ongoing UNDP Philippine projects.

Finally, we also have access to the "Project Center Sharepoint", the "Project Document Center", and the "UNDP Philippines Portfolio Review Dashboard", which helped the cross-validation of the data and provide the baseline for the current analytical efforts and resources.

[5] https://undp.gitbook.io/pasi/

Data Model

The data model is what allows us to organize and relate to the data that we gather during the project. As a result, it is not a static object but rather a model that evolves and responds to the ingestion of new data and the elicitation of new requirements or questions.

The figure below presents a simplified version of the working data model. Each node is an entity in the knowledge graph and each connection represents a relationship between a pair of entities.

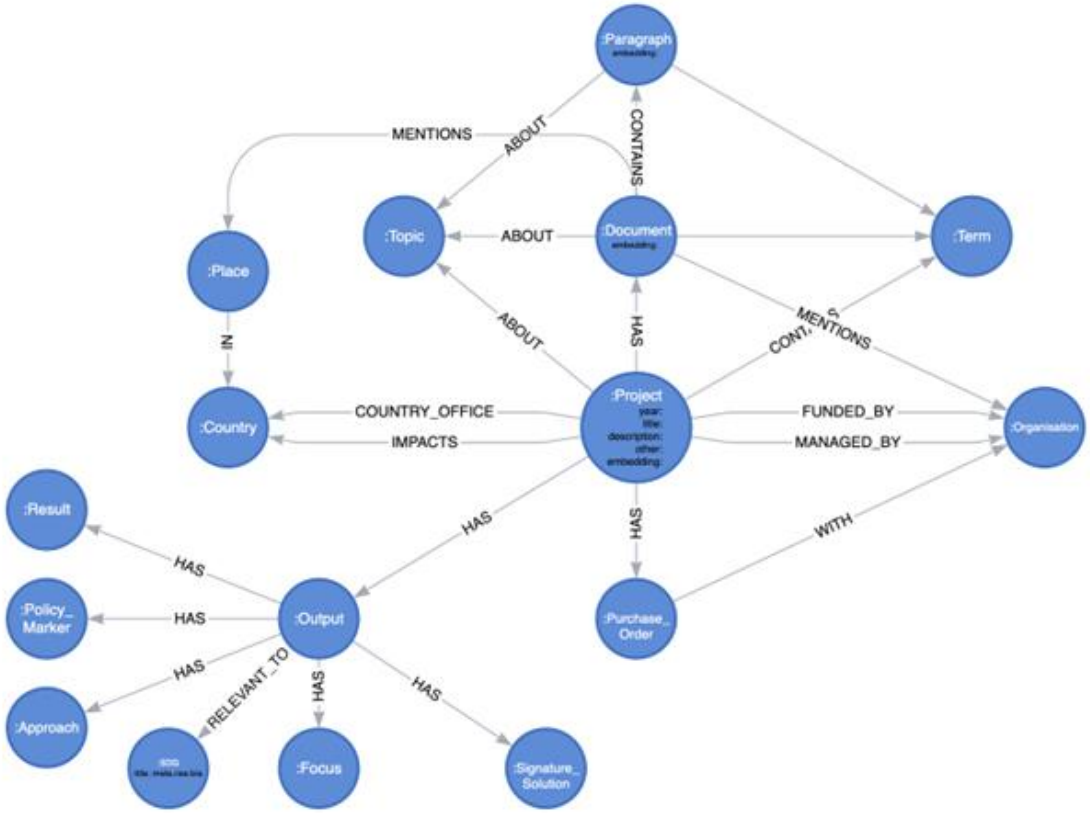

Figure 8: Simplified version of the working data model.

Computational Methods

To create a knowledge graph that enables new project portfolio insights, we followed the traditional ETL pipeline of data Extraction, data Transformation and data Loading into a graph and later custom dashboards that leverage tools such as Tableau and Graphext.

Here, the main departure from standard ETL is the addition of more advanced methods in the transformation and loading steps. Although a commonly used umbrella term for this kind of methods is "AI", to be more precise, we can divide the two most important computational methods used in this project into two:

1) **Network Analysis Methods:** These methods are what allow us to quantify structures within the knowledge graph such as clusters as well as the relative distances/similarities between the clusters and individual nodes, places, or other entities. The results of the application of these methods are most visible in the network graphs and the clusters identified in the network graphs.
2) **Data Labelling Methods** (including natural language processing)**:** In layman terms, these are the methods that allow us to identify/extract a subject or assign a topic/classification/category/etc. to a piece of text, such a section within a project document or project description. These methods are a key part of the sensemaking efforts because they seek to scale up the recognition of patterns that help make sense of a portfolio of projects. The application of these methods was instrumental to for example label/classify the projects according to UNDPs core offerings.[6]

The integration of these methods with network analysis helps us to not only see individual projects and their characteristics but also see the connections between the identified characteristics and highlight, for example, potential collaboration areas between the projects.

One of the more challenging and distinct parts of the analytical work was to develop a custom data labelling method that could classify projects into custom classes/themes/topics having only one single description for each class/theme/topic. To do this, we combined pre-trained topic classifiers with entity extraction and recognition, graph embeddings and language models.

The figure below provides a simplified diagram with the applied custom method:

[6] https://www.undp.org/expertise

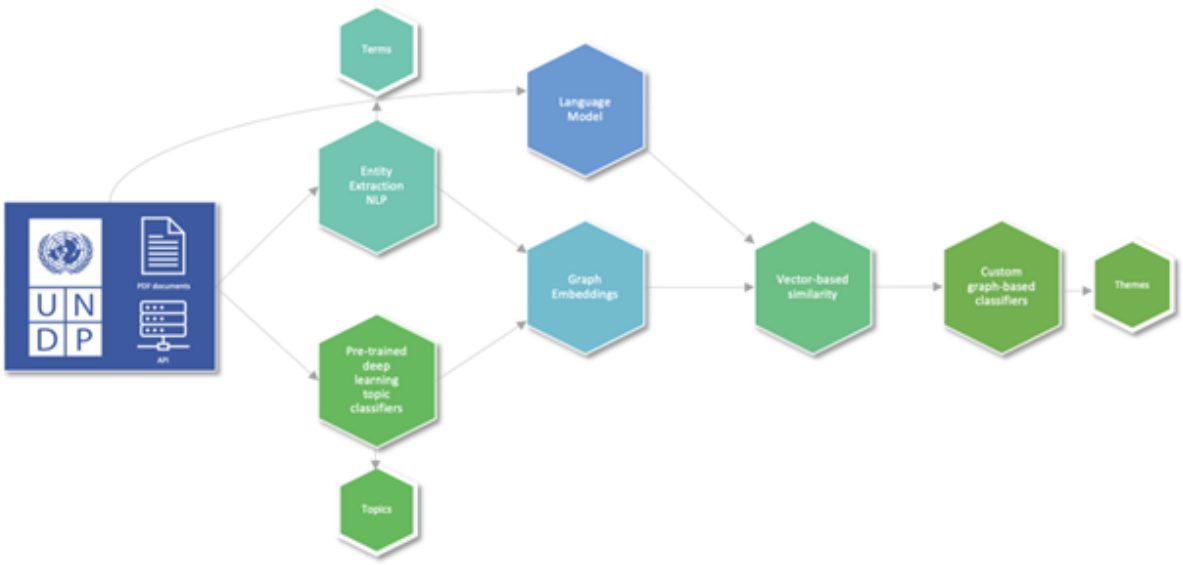

Figure 9: The applied custom method[7]

Architecture and IT Infrastructure

Solution overview

The proposed solution can be divided into four core modules, three on the backend and one on the front-end:

Back-end:

- **Data ingestion and preparation module:** Allows ingesting the set of PDF documents and metadata, storing them into a unified document database. This also enables us to normalize the metadata across the different documents.

[7] The following research papers provide a more detailed technical reference about the different steps followed to derive the custom graph-based classifiers

Entity Extraction NLP:
Al-Moslmi, Tareq, et al. "Named entity extraction for knowledge graphs: A literature overview." IEEE Access 8 (2020): 32862-32881.

Pre-trained deep learning topic classifiers:
Shen, Zhihong, Hao Ma, and Kuansan Wang. "A web-scale system for scientific knowledge exploration." arXiv preprint arXiv:1805.12216 (2018).

Language Model:
Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Graph Embeddings:
Goyal, Palash, and Emilio Ferrara. "Graph embedding techniques, applications, and performance: A survey." Knowledge-Based Systems 151 (2018): 78-94.

Vector-based similarity:
Xia, Peipei, Li Zhang, and Fanzhang Li. "Learning similarity with cosine similarity ensemble." Information Sciences 307 (2015): 39-52.

Custom graph-based classifiers:
Guo, Qingyu, et al. "A survey on knowledge graph-based recommender systems." IEEE Transactions on Knowledge and Data Engineering (2020).

- **Data linkage module:** this module creates a shared relational structure between the documents and loads the structure into a graph database that can later be used to store the enriched documents and to connect them into project portfolios.
- **Data management, enrichment, and analysis module:** this module allows to further normalize, enrich, and link data using natural language processing (e.g., BERT and entity extraction), network analytics to embed new metrics and indicators services and provides graph analytic capabilities. This module also enables backend data management and supports an input/output API layer.

Front-end:

- **Information visualization module:** This module provides the tables and charts exposed to internal and external users according to the filters and parameters sets for each query.

In addition to the modules described above, we built in the system components to monitor server errors and problems, as well as means to periodically backup both the data and the platform itself. The next figure provides a visual high-level summary of the modules previously described.

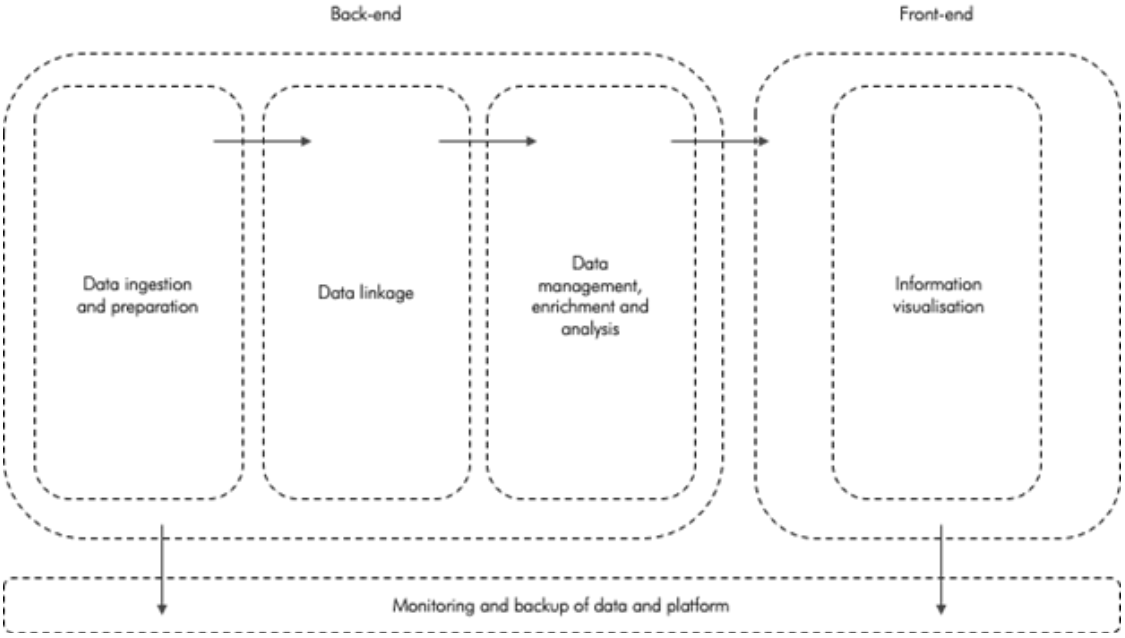

Figure 10: High-level summary of the overall modules and their relations

Architecture Overview

Following the description of the software components previously introduced, we developed a modular containerized architecture where key components and modules are packaged within Docker containers with each Docker connected with the rest through transparent APIs. The figure below provides a high-level overview of the system architecture and its key components.

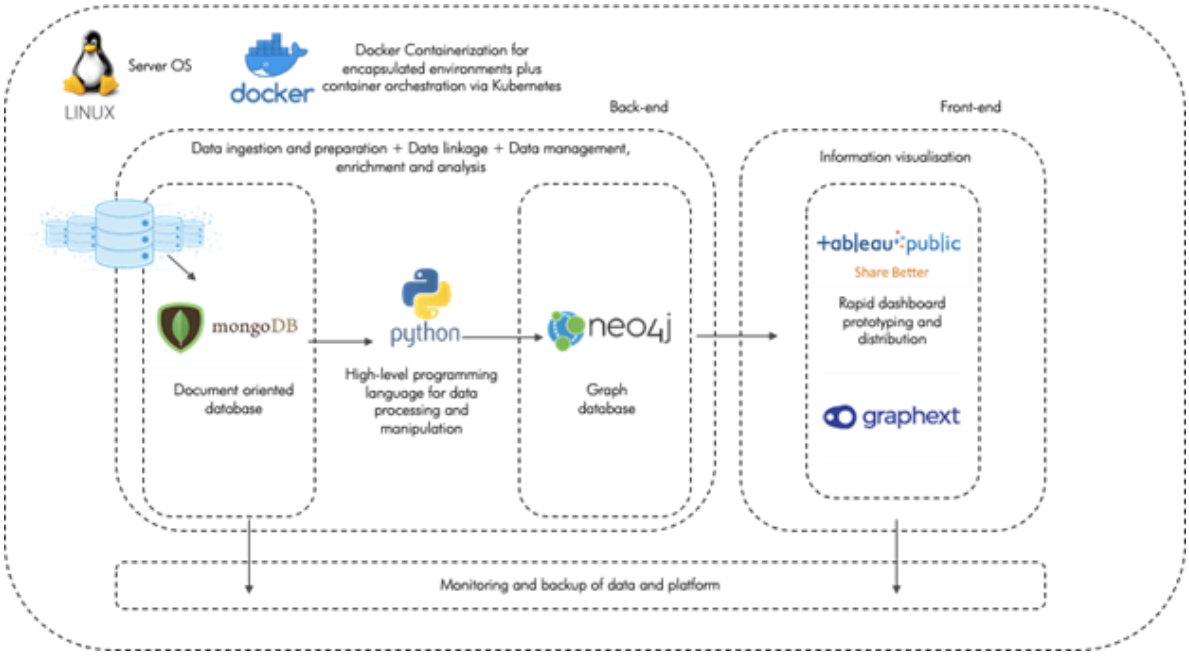

Figure 11: Overview of the system architecture and its components

Guiding Principles

In this project we followed the following set of principles to define the technical architecture of the system and to select the software components that form part of the solution:

System modularity	Cross platform and component interoperability	Flexibility and reusability of components and functions
Preference for open source components	Built in scalability	Transparency (for machine and human)
Wide availability of community support for selected components	Vendor independence (for hardware and software)	Self-administration of the platform by default
Robustness to obsolescence	Human centric usability	Cost efficiency and effectiveness

Guiding Principles for the Technical Proposal

Figure 12: Guiding principles for the technical proposal

Software Components

The software components can be divided into 1) server OS and containers, 2) development libraries and frameworks for the front-end and the back end of the system, 3) database components, and 4) packaged applications.

The following list details the core software components for each category:

Server OS and containers:

We hosted the services on a Linux environment using Docker as our container ecosystem to package and run the developed modules.

Development libraries and frameworks:

Back-end:

- We used Python as the core high-level language and Python libraries such as Pandas and Numpy for core data analysis and manipulation activities. In addition, for PDF work we incorporate Apache PDFBox, for NLP analytics Google´s BERT, custom entity extraction pipelines and machine learning classification via Scikit-Learn and Pandas.

Front-end:

- We used D3.js, Tableau for rapid dashboard prototyping as well as Graphext and React components as some of the core tools to develop the required web functionalities.

Database components:

- We use MongoDB as the core Document Database and Neo4j as the core Graph Database. For these two databases we use their respective open source/community versions.

Packaged applications:

- We support and implement the connection with Tableau and provide functionalities to export tabular files that can be imported into statistical packages such as SAS, SPSS and Stata.

Hardware Components

During the development process we deployed the different components and services in Dataverz servers. Afterwards we installed within the UNDP key components of the infrastructure to allow self-hosting of the data and for running updates using the support of UNV.