Humanitarian Data Insights Project - FAQs

Can we use other public humanitarian datasets in this tool? (If so - how?)

Yes, it has been designed so that APIs that adhere to the openapi standard can be added through configuration. More complex APIs can also be added and custom code developed if organizations pref.

Why did we choose the HAPI data source / what does that unlock for us?

We have worked extensively with the HDX team and the new HAPI API is a great new product for normalized Humanitarian data. Given these two factors, we were delighted to use HAPI as one of our first data sources. Also, being such a nice API, it's the perfect candidate to test a new approach such as data recipes as it opens up a wealth of humanitarian data in a standardized format.

How real-time is the tool/data expected to be?

For data sources which are configured as API data sources, the system will call them on-demand, so it will pull in the latest data on the remote system [Note: The demo env doesn't have this configured]

For data sources where we ingest data, like HAPI, it depends how often that ingestion is run. That is controlled by the user of the humanitarian ai assistant

What are the use cases that this tool performs well for / likely performs well for?

Humanitarian AI assistant and its underlying architecture ‘Data recipes’ is targeted at individuals and organizations that want to get insights from humanitarian data to help them in their work, with less effort, cost, and without having to have advanced data analytics skills.

The underlying architecture for the humanitarian AI assistant is a product DataKind have developed called ‘Data Recipes AI’. This architecture save data analysis ‘Recipes’ to a database, programmed with AI assistance, where they can be reviewed by a human and used on websites, search engines and chatbots. So the architecture has many potential applications across any domain with data.

Data Recipes AI performs best when the data it’s using can be easily extracted, with via an API on the fly, or ingested, or in a database. If however the APIs are very complicated, or don’t offer a good way to filter data, then that will be a challenge as the system is limited to the interfaces being provided.

There are no doubt other challenges we will face down the road, but being so new, we are excited when launched to get feedback from the community in order to tackle these.

How does this tool work in low bandwidth environments?

We have designed recipes so any large data transfer and processing is done centrally. People using the chat interface, for example on their phone, don’t need to transfer a lot of data, it’s images of graphs and snippets of text.

It would also be possible to host the central data processing in a low bandwidth environment if the data is physically located there. The solution can run on just local data, so if somebody installed that locally, for example a dump of the Data Recipes DB after ingesting HAPI data, then the tool could be run locally with low bandwidth.

It is worth noting, that right now we are using LLMs in the cloud. The centralized nature of data recipes means this traffic is intentionally limited for chat users, all the heavy lifting (prompting)is done in the central location, but it does currently require an internet connection. That said, we have implemented the LLM part using a framework called LangChain, which can also use LLMs running locally. So on ut roadmap, it would be possible run everything locally, assuming you have the data, with no internet connection.

What LLMs does the conversation data analysis tools use? Does it have to use only OpenAI LLMs like GPT 3.5 and GPT 4?

For our initial work, we are using OpenAI models, but we have developed the code to use the Langchain framework so that any of its supported models - including open models - can be configured.

We have also intentionally designed the system such that the model used for end-users doesn't have to be very powerful. This is the data recipes concept, part of which is aimed at reducing costs.

Product description - Humanitarian AI Assistant

"Humanitarian AI assistant and its underlying architecture 'Data Recipes' is targeted at individuals and organizations that want to get insights from their data to help them in their work, with less effort, cost, and without having to have advanced data analytics skills."

- **Purpose:** Give people a more user-friendly way to find and use data for their work
- **Benefits:** Reduces the effort required in generating reports and processing data needed as part of work such as humanitarian response, in some cases opening access to new insights that were not previously available. For data analytics teams, reduces the overhead in providing data products such as graphs and reports by helping them build a reusable library assisted by AI, which their end users can then use on demand
- **Features:** Intuitive conversational interface; AI-assisted context-aware programming for creating new graphs and reports which reduces the time needed to create new assets; A way to maintain a library of data analysis tasks (recipes), such as creating graphs and text; A human review workflow; Trust features to indicate is results are from assets approved by a human, versus on-the-fly LLM analysis; Possibility of a community hub of data recipes
- **Differentiation:** There are many LLM-assisted data analysis products available, but to my knowledge, none blend LLM-generated and human-generated tools in the same way, or offer a badge system to indicate if the LLM response has been human-reviewed. Though

initiatives like Google Data Commons offer a set of standard data visualizations, I don't believe they support conversational generation of community input.

How are you managing LLM safety?

Primarily by design. The concept of data recipes have a human review at its core before LLM-generated assets (ie the code to produce graphs and reports) are presented to end users. In this way many of the safety issues are mitigated. It is possible also have LLM-generated analysis in the system, but we have implemented a badge system to inform the user what has had human review and what has not. Finally, we use LLMs in cloud services which implement real-time content safety monitoring.

How is using the recipes CLI to generate and edit recipes different to GitHub Copilot?

Generating and editing recipes by prompting an LLM is similar to Copilot's review of code, and that's the intention because we felt data scientists will be familiar with LLMs in their workflow. The key difference is that the prompt is dynamic and automatic, informing the LLM about data within data recipes and available previous recipes so it can learn. It also fits in with the recipes publishing workflow, will rerun code to debug and fix it and can communicate with external sources, all things Copilot cannot currently do.

How do I get the Humanitarian AI assistant?

Right now we are working with a small working group, who are checking out the solution from github and running locally. That is helping us simplify the installation process as well as refine the solution in general. If you would like to be part of that, please contact us.

We are aiming for full release later in the summer.

What is the technical architecture?

Data recipes AI runs as a docker application, with components for ingestion, data storage, chat interface and AI-assisted command line interface for creating recipes.

Won't there be overhead for the recipe managers maintaining and approving data recipes?

Though AI data analysis on-the-fly is no doubt useful, we all know if can sometimes fail or produce incorrect, but convincing results. We felt that combining this ability with reviewed human-approved recipes - with clear indication to alert the user about how their analysis is being done - would provide a more stable offering and build trust.

So yes, there is overhead in managing recipes, but we have provided AI tools to try and make that more straightforward where an analyst can conversationally program to create recipes. Also, as the recipe library grows - even with community input, image key advanced visuals for HAPI which everyone can use - then generating new recipes may well become easier over time as the AI can learn from human-approved work.

How will the system map ambiguous placenames in the data?

Currently, place names are found by looking up names in the HDX data extracted from the new HAPI. We rely on the LLM to extract this data, and map to the HDX data. In the next few weeks we are hoping to implement a vector search across the HAPI data in the database for more nuanced entity matching of places.

==== *from user Community of Practice, in quick note form, added by Rachel* ====

Seems relevant across sectors, does DataKind plan to release open source and/or use across other sectors?
Yes! Already thinking about other projects internally at DK where it is applicable

Anyone can add data to HDX: what does governance/legality for data across countries look like?
HAPI datasets are selected by HDX team as reliable public datasets

What is our requirement internally - what are we responsible for vs. HDX vs. DataKind - for data legal alignment?
Responsibilities are divided / it depends on role:

- Data uploaders and HDX responsible for data governance for data added to HDX
- DataKind responsible for ensuring tool is built such that internal data only lives within environment is being run (ie we won't use your data to train/inform the tool being used at a different organization)
- SCI responsible for ensuring tool is deployed within their own secure infrastructure if they'd like to use their own secure data

Deployment with your own data internally? How secure is that?
Yes, it would all exist within your organization's infrastructure, it would depend on security of your infrastructure.

In the long term when this is ready, you don’t have to add any recipes or data to cloud for other orgs to access. Model will not use internal data to train on other organizations.

Can we share with other orgs for relevant resources?
Yes - this is part of the vision for the open source tool and community collaboration

Is this hosted in your own environment?
DataKind is hosting our own, SCI wants to host their own and run locally, you can do either

Offering for providing technical feedback
Need to identify the best places and approach for this - but yes - excited to get your insight!