Implementation Network for Sharing Population Information from Research Entities in East Africa

(INSPIRE - EA)

INSPIRE EA IT Infrastructure on the Cloud

Version 1.0

Version 1.0

Version	Date	Author	Reviewer	Description
1.0	2021-07-14	Tathagata Bhattacharjee		This version outlines the processes involved in setting up the IT infrastructure on Cloud for INSPIRE EA.

ACKNOWLEDGEMENT

This initiative was tasked with setting up an IT infrastructure on the cloud for hosting INSPIRE EA data hub to store and process data from the collaborating institutions. This project was undertaken as a part of the project funded by the UK Research and Innovation's (UKRI) Global Challenges Research Fund (GCRF) Digital Innovation for Development in Africa (DIDA).

Reference No.: EP/T029315/1

Project Title: Implementation Network for Sharing Population Information from Research Entities in East Africa (INSPIRE-EA)

Table of Contents

Introduction

This document introduces the processes involved in setting up the IT infrastructure on the cloud for hosting the data hub of INSPIRE EA network. The main objective here is to outline the approaches, methods, and implementation tasks that were undertaken to set up this infrastructure to facilitate data storage, data processing, and data accessibility features for the network.

This setup allowed the storage of input datasets in various formats along with applications for extracting, transforming, and loading pipelines to populate the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). It also provided access to tools for data mapping, data documentation, and data analysis for respective roles within the collaborating and network partner institutions.

Selection of the Server Location

To host the data hub for INSPIRE EA it is necessary to establish a common server to service all requirements of data storage, processing, sharing, analysis etc. The IT infrastructure was conceptualized in such a way that all stakeholders, including the collaborating and partner institutions along with the data end-users, must have seamless and flawless access to the server, with fine grained access control to achieve their respective role goals.

The initial discussion revolved around setting up a server that would be optimally configured for the purposes mandated for the success of INSPIRE EA implementation. Having a server setup within a collaborating institution's data center was the first idea that was discussed. This type of implementation had the advantages of holding the infrastructure within the network's physical boundaries. This thought process was born from the General Data Protection Regulation (GDPR) guidelines, which encourages data storage in the country or the region where it was generated. Due to the upcoming and yet unclear data access regulations differing from country to country, a setup within the network's institution was thought to be the best option and technical feasibility was sought from the partner institutions. The response from the institutions marked some common difficulties, such as acquiring a space within the respective institution's data centers (the IT administration team had many queries and limitations in providing access to networks outside), internet bandwidth limitations, and shortage of technical human resource to manage such setups.

To have a server setup that required less maintenance overheads while simultaneously being accessible, secure, and available 24/7, the technical team of INSPIRE EA did some brainstorming and decided to go for a cloud service provider to host the server.

Cloud Services

With the exponential increase in data usage in every vertical of businesses and research, it has become increasingly difficult for individuals and organizations to keep all important information, programs, and systems running on internal computer servers. Cloud services are services that are available through remote cloud computing server(s) rather than on-premise server(s). The cloud services provide scalable solutions that are managed by the service provider and provide users with access to computing services with benefits of more productivity and improved efficiency to significant cost reduction and simplified IT infrastructure management.

The benefits of cloud services are listed very briefly here:

Cost benefits: Cloud services significantly reduce the IT costs by eliminating the need for IT hardware and system administrators. Cloud services are a good option for small organizations, startups, research groups for they can readily start on the core work rather than on setting up the IT infrastructure themselves.

Better security: Implementing security on private clouds is significantly easier than relying on hardware and skilled security experts within the organization. Unauthorized access to the data on cloud is much rarer because of the security implementations on cloud services for compliance and business sustainability reasons. Backup, encryption and data recovery are much more streamlined compared to the traditional methods.

Handling spikes and scalability: The first advantage that comes to an IT architect's mind in opting for a cloud service is the ease of the scalability feature. This means that we can increase or decrease the size of the servers depending on demand without affecting the performance or the end-user experience. The occasional increase of loads of data storage, processing, or an increase in number of users are called spikes. The cloud services can be scaled up during these spike cases and fall back to the standard configuration at other times. This helps to save costs through the year as we do not need to maintain the peak load configuration all the time.

Backup and Recovery: Many costs as well as human resource allocation can be saved on infrastructure and maintenance by hosting data storage and applications on the cloud. The cloud service provider will be responsible for the data security and compliance matters. It increases the flexibility with on-demand backup, large storage capacity, and faster restoration in cases of failures.

The top cloud service providers globally in the year 2020 were:

1. **Amazon Web Services (AWS)** is the market leader which operates in 20 geographical regions across the world and offers 175 fully-featured services to meet any kind of server services requirements.
2. **Microsoft Azure** is the fastest growing cloud giving a tough competition to AWS and other cloud service providers. Azure has 54 data center regions across the world with availability in 140 countries.
3. **Google Cloud Platform (GPC)** is also a competitor to both AWS and Azure and excels in providing services with least latency for high performance-oriented applications. GPC is available in 22 regions, 61 zones and 200+ countries.
4. **IBM Cloud** offers a host of IaaS, SaaS and PaaS services via public, private, hybrid and multi-cloud models with around 170 products and services ranging from Internet of Things (IoT), Cognitive Computing and Blockchain.
5. **Oracle Cloud** is an Enterprise Resource Planning (ERP) based cloud service to build, deploy and manage workloads in the cloud or on-premise.

There are many more cloud service providers and each has its own strengths and relative areas of use.

Selection of the Cloud Service Provider

Once the decision was made that a cloud service provider would be used to host the INSPIRE EA data hub, the next step was to select the one that would be most suitable for our type of requirements. Our requirements were simple and are listed here:

1. The cloud service provider must have its data center in the African continent. Ideally, the data must be hosted within the country where it was generated but since INSPIRE is a network of institutions from many countries in Africa, having the option of having the data center of the cloud service provider in each of the countries was an absolute negative proposition. Thus, we decided to look for cloud service providers with data centers in Africa. One of the guiding principles for making this decision was the General Data Protection Regulation (GDPR) guidelines.
2. The INSPIRE team members' experience working with cloud services guided us to Amazon Web Services and Microsoft Azure.
3. After researching these two cloud service providers, we realized that both were very compatible with our requirements and would provide similar services within the context we were looking for. The AWS had a slight higher edge over Azure asit inherently supported the creation of data science environments for health analysis using OHDSI (OHDSI-on-AWS: https://aws.amazon.com/blogs/big-data/creating-data-science-environments-on-aws-for-health-analysis-using-ohdsi).
4. Our hopes of continuing with AWS soon vanished when we realized that it did not have a data center within the continent of Africa, at that point in time. Microsoft Azure, on the other hand, had a few data centers in the South Africa region, which made it compliant with our requirements. Thus the decision was made to opt for Azure with a data center in South Africa.
5. The next step was registering with Azure and making payments for the services.
6. After registering on the Azure portal (https://portal.azure.com/) , we started looking for data centers and found that Azure has two centers in the South Africa region, and we decided to move forward with the project.
7. Now, the next was to register the payment option and it was found that Azure accepts credit cards as the only option. The credit card to be used was that of a personal card of PI Jim Todd, and unfortunately Tanzania does not offer credit cards, so he used his UK credit card.
8. Now that the payment method was registered with a UK credit card, we were not able to launch the Africa region virtual instance but had to go with regional data centers from either the United States or Europe..
9. Eventually we opted for the UK region and launched the virtual instance in the UK South region. However the services are transferable and we hope to locate the service in Africa once we can establish a sustainable payment method through the INSPIRE Secretariat.

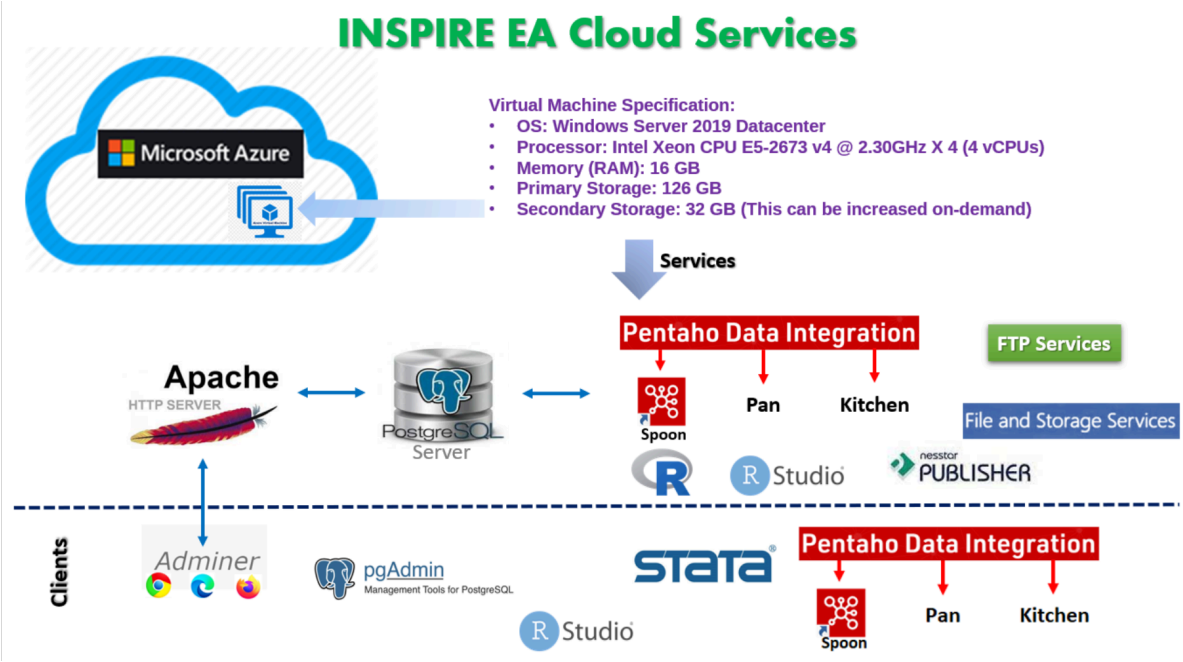
Selection of the Virtual Server (Virtual Machine)

The virtual server that was planned to host all the services needed to be on a platform that would be familiar to the INSPIRE EA technical team members. The options that Microsoft Azure provided were Windows server and Linux Ubuntu Server. Since, all members were not very conversant with the Linux environment, it was decided to go with the Windows server instance. It was also decided that all the

applications and tools would be installed in the server instance itself and Azure resources, like the Web App and Database, would not be configured.

INSPIRE EA Virtual Machine on Azure Cloud

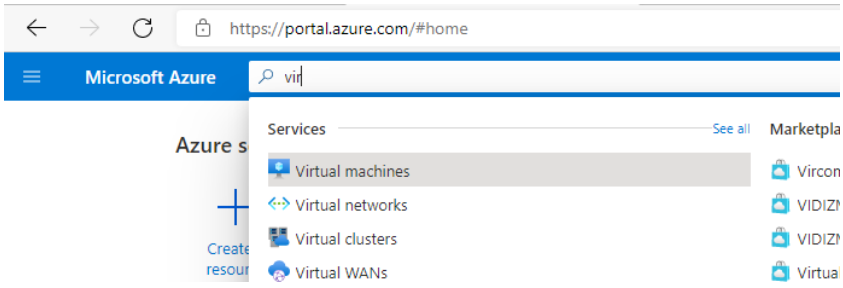
The high-level layout of components of INSPIRE EA virtual machine in Azure cloud is shown in the block-diagram below. The Virtual Machine (VM) has a set of host-based applications and web services. Each of the installed applications are explained in the sections later in this document.



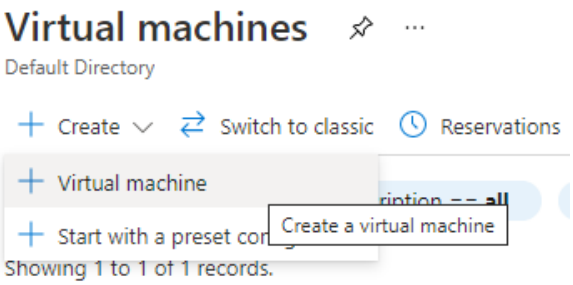
Windows Virtual Machine in Azure Cloud

Azure virtual machine was created through the Azure portal using the browser-based user interface.

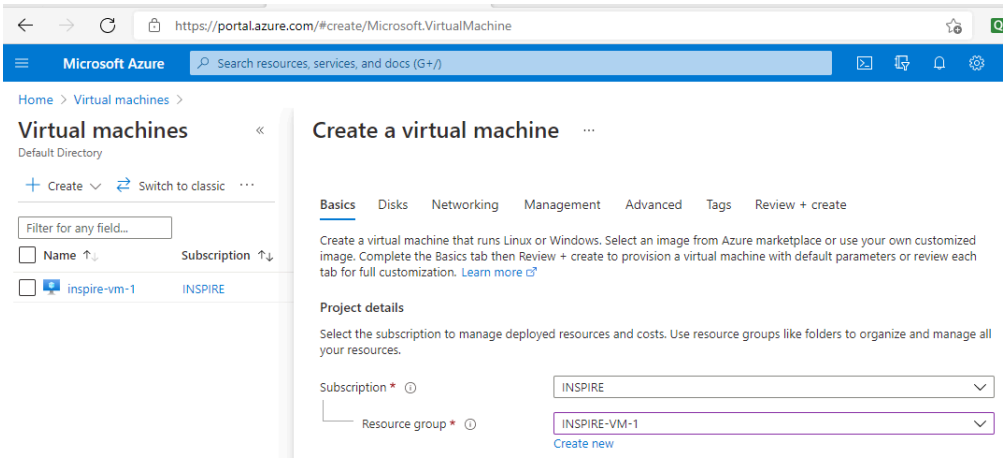
1. Signed in to the Azure portal: https://portal.azure.com/
2. Typed **virtual machines** in the search



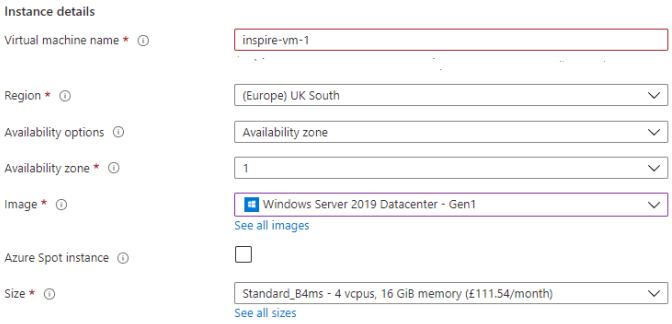
3. Under **Create**, selected **Virtual machine**



4. In the **Basics** tab, under **Project details**, selected the subscription **INSPIRE** (our Azure subscription name which was created earlier as pay-as-you-go). Next created a new resource group named INSPIRE-VM-1 by clicking on the **Create new** link.



5. Under **Instance details** tab, for the **Virtual machine name**, we named it inspire-vm-1 and choose (Europe) UK South for the **Region**. We then chose Windows server 2019 Datacenter for the **Image** and Standard_B4ms – 4 vcpus, 16 GiB memory for the **Size**.



6. Kept the remaining tabs to its default values and clicked on the create button to create the instance. The details of the instance are shown in the screenshot below.

Virtual machine

Computer name	inspire-vm-1
Operating system	Windows (Windows Server 2019 Datacenter)
Publisher	MicrosoftWindowsServer
Offer	WindowsServer
Plan	2019-Datacenter
VM generation	V1
Agent status	Ready
Agent version	2.7.41491.1010
Host group	None
Host	-
Proximity placement group	-
Colocation status	N/A

Availability + scaling

Availability zone	-
Scale Set	-

Networking

Public IP address	51.105.33.160
Public IP address (IPv6)	-
Private IP address	10.0.0.4
Private IP address (IPv6)	-
Virtual network/subnet	INSPIRE-VM-1-vnet/default
DNS name	Configure

Size

Size	Standard B4ms
vCPUs	4
RAM	16 GiB

Disk

OS disk	inspire-vm-1_OsDisk_1_ceba4c87862449a8a7552007ed932b3a
Azure disk encryption	Not enabled
Ephemeral OS disk	N/A
Data disks	0

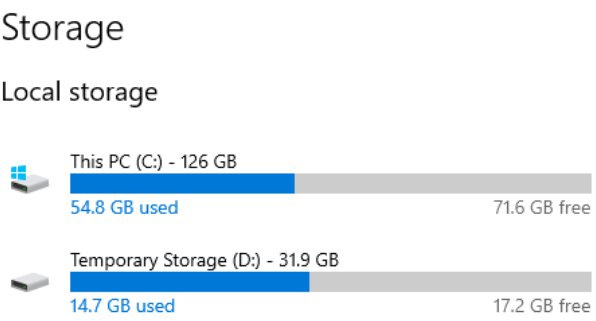
7. The device details are as follows:

Device specifications

Device name	inspire-vm-1
Processor	Intel(R) Xeon(R) CPU E5-2673 v4 @ 2.30GHz 2.29 GHz
Installed RAM	16.0 GB
Device ID	99875357-3051-4198-B233-F6430DED3ECF
Product ID	00430-00000-00000-AA707
System type	64-bit operating system, x64-based processor
Pen and touch	Pen and touch support with 10 touch points

Storage details

Since this was a pilot activity under the INSPIRE EA phase I implementation, we have not taken any additional storage. We had a primary drive of 126 GB with a secondary temporary storage drive of 32 GB, which is default under the VM B4ms Azure image.



Network Security Group

A network security group contains security rules that allow or deny inbound network traffic to, or outbound network traffic from, several types of Azure resources.

Here we have created a network security group and have named it inspire-vm-1-nsg. The details of this are shown in the screenshot below.

Resource group (change) : INSPIRE-VM-1
Location : UK South
Subscription (change) : INSPIRE
Subscription ID :
Tags (change) : INSPIRE-Virtual-Machine-1 :

Custom security rules : 5 inbound, 0 outbound
Associated with : 0 subnets, 1 network interfaces

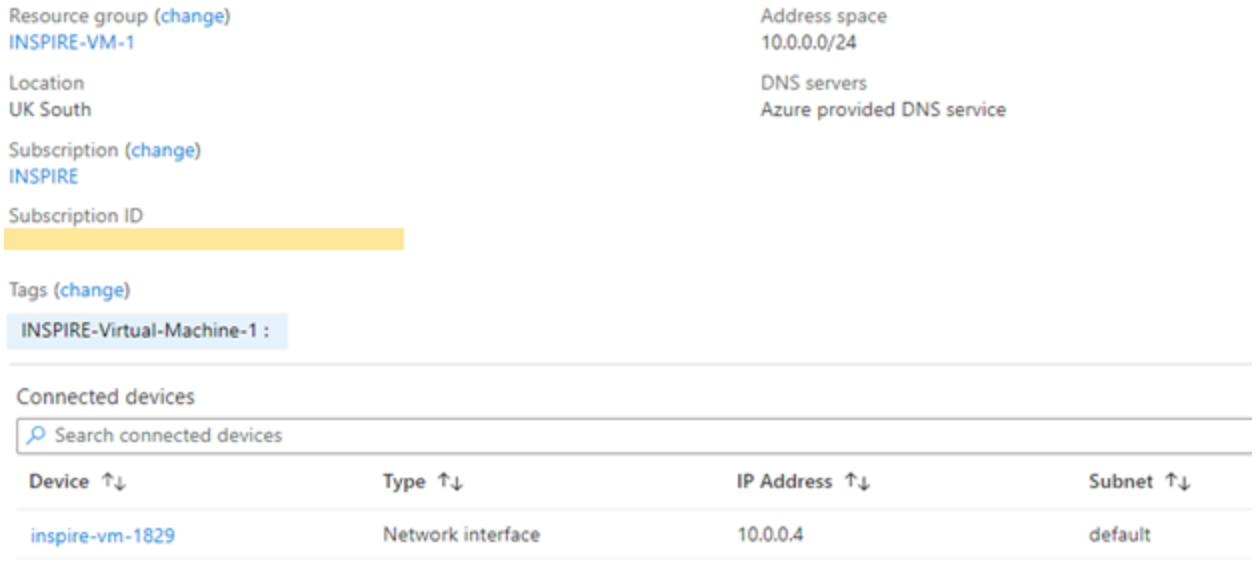
Filter by name | Port == all | Protocol == all | Source == all | Destination == all | Action == all

Priority	Name	Port	Protocol	Source	Destination	Action
Inbound Security Rules						
300	RDP	3389	TCP	Any	Any	Allow
320	HTTP	80	TCP	Any	Any	Allow
340	HTTPS	443	TCP	Any	Any	Allow
350	PostgreSQL	5432	TCP	Any	Any	Allow
360	IISWebServer	8080	TCP	Any	Any	Allow
65000	AllowVnetInBound	Any	Any	VirtualNetwork	VirtualNetwork	Allow
65001	AllowAzureLoadBalancerInBound	Any	Any	AzureLoadBalancer	Any	Allow
65500	DenyAllInBound	Any	Any	Any	Any	Deny
Outbound Security Rules						
65000	AllowVnetOutBound	Any	Any	VirtualNetwork	VirtualNetwork	Allow
65001	AllowInternetOutBound	Any	Any	Any	Internet	Allow
65500	DenyAllOutBound	Any	Any	Any	Any	Deny

Virtual Network

The Azure Virtual Network is a logical representation of the network in the cloud. So, by creating an Azure Virtual Network, we defined our private IP address range on Azure to deploy different kinds of

Azure resources. As of now, we haven't created any additional Azure resources so effectively this virtual network is only connecting the virtual machine. The details of this are shown in the screenshot below.



Public IP

A public Internet Protocol (IP) address is the address that is assigned to a computing device to allow direct access over the Internet. We needed a public IP address to access the services hosted on the virtual machine from over the internet. The details of the IP address are shown in the screenshot below.

Network Interface

A network interface enables an Azure VM to communicate with the internet, Azure, and on-premises resources. A VM has one or more network interfaces. Here, we have used the default Network Interface Card (NIC) that gets created along with the creation of the VM. The details of the network interface in our VM are shown in the screenshot below.

Resource group (change)
INSPIRE-VM-1

Location
UK South

Subscription (change)
INSPIRE

Subscription ID
[redacted]

Accelerated networking
Disabled

Private IP address
10.0.0.4

Public IP address
51.105.33.160 (inspire-vm-1-ip)

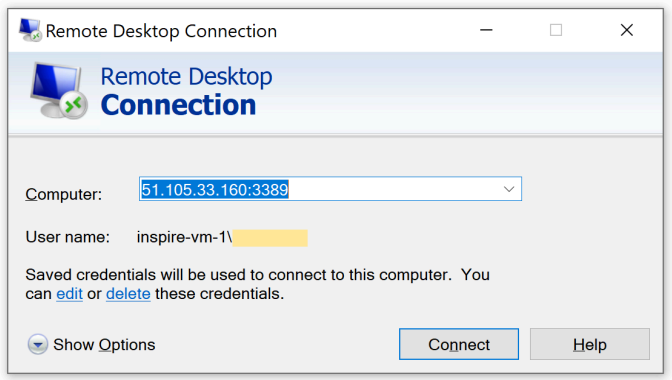
Private IP address (IPv6)
-

Public IP address (IPv6)
-

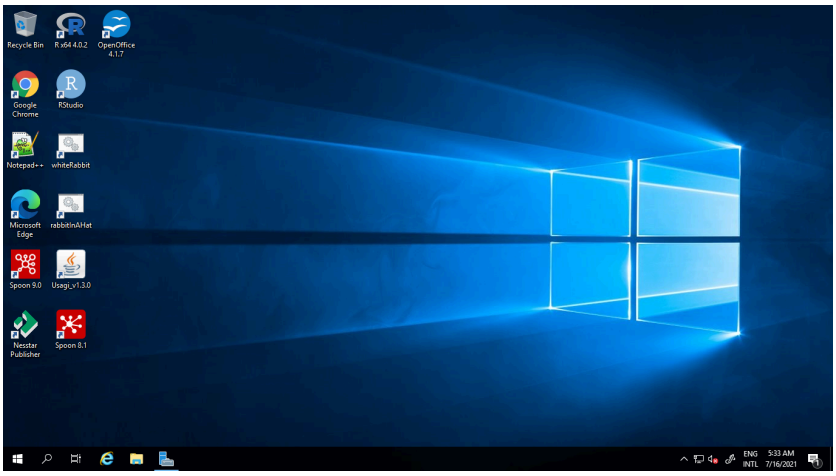
Attached to
inspire-vm-1

Connection to the Virtual Machine using Remote Desktop Connection

To connect to the remote virtual machine on Azure, we use the window's Remote Desktop Connection app.



After successful connection, we get into the Windows Server 2019 Datacenter remote virtual machine.

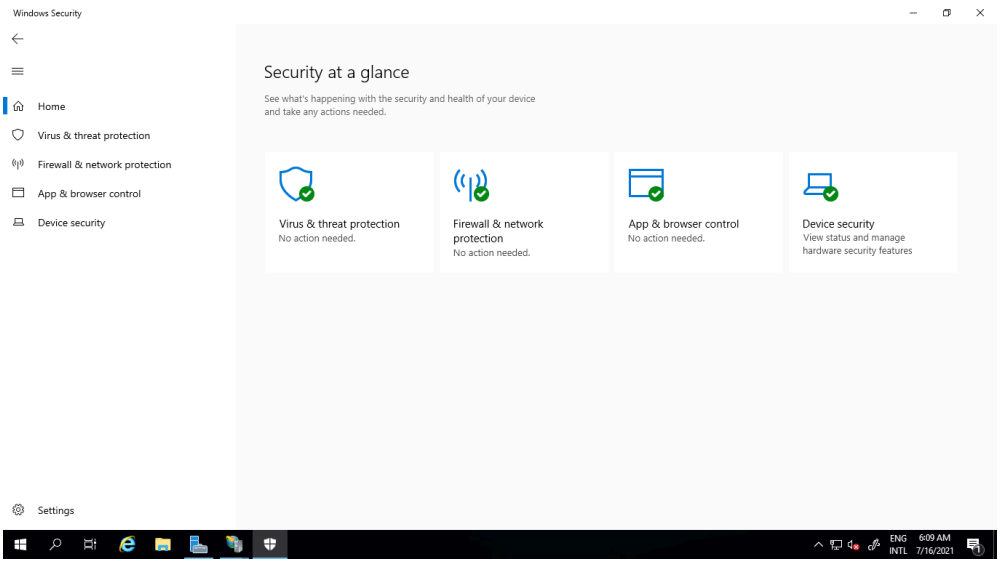


Multiple Remote Desktop Connections on Windows Server 2019

By default, Windows Server 2019 / 2016 / 2012 allows only a single Remote Desktop session. We needed multiple users to work on the server at the same time for various data documentation, data mapping, data analytics programming, etc. Our Windows Server 2019 on Azure VM has been enabled to support multiple remote desktop connections and the limit is set to maximum 5 which can be increased if needed.

Windows Security

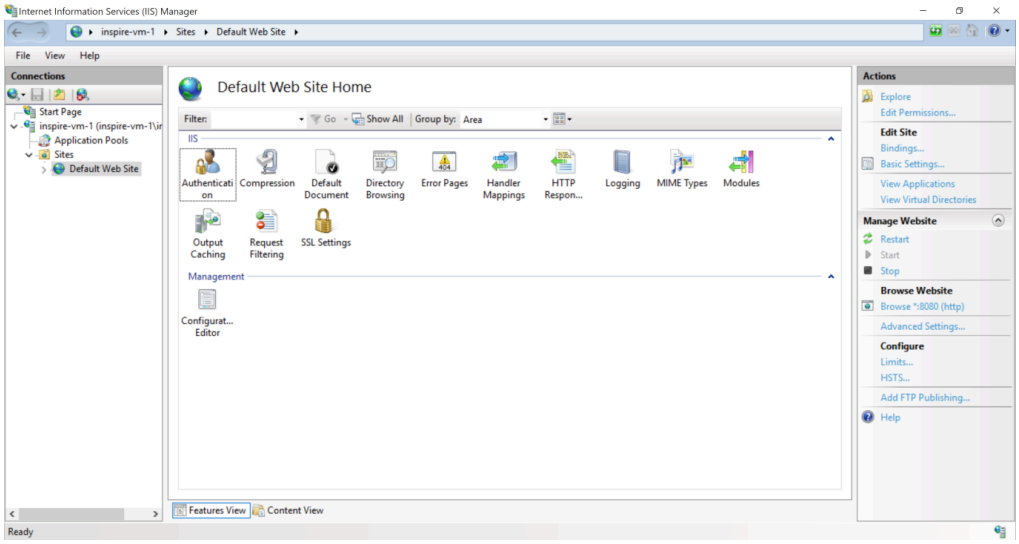
We have enabled the default Windows Server 2019 security features, which provides comprehensive security to the Windows virtual machine. At a glance, the security features that are enabled are: Virus & threat protection, Firewall & network protection, App & browser control, and Device security. The following screenshot shows the security at a glance.



Internet Information Services (IIS) on Azure VM

Internet Information Services (IIS) is an extensible web server software from Microsoft, which accepts and responds to the client's computer requests and enables them to share and deliver information across the internet. The web services on our VM are hosted on IIS that can be accessed by our users with client web browsers like Microsoft Edge, Google Chrome, Mozilla Firefox, etc.

The following screenshot shows the IIS web server configured to host the default website of the data hub.



Apache Web Server

The Apache HTTP Server is a free and open-source, cross-platform web server software that was released under the terms of Apache License 2.0. Apache is developed and maintained by an open community of developers under the auspices of the Apache Software. It is a robust, commercial-grade, feature-rich, and freely available source code implementation of an HTTP (Web) server.

We have installed and configured our system to run Apache 2.4 on Windows. The source Apache 2.4 binaries were downloaded from Apache Lounge (https://www.apachelounge.com/download/). The purpose of installing Apache was to host the open-source web tools from OHDSI and other providers.

PHP

PHP is a server-side scripting language for building dynamic and interactive web pages. It is a widely-used open-source scripting language. We installed PHP 7.

Many web applications use PHP and thus for enabling any such application on our data hub, we installed PHP. One such use was for the data access user-interface named Adminer, which needed PHP to run and provided web-based access to our PostgreSQL database.

SSL Certificate

Secure Sockets Layer (SSL), is a computing protocol that ensures the security of data sent via the internet. This protocol is for web browsers and servers that allows for the authentication, encryption, and decryption of data sent over the internet. When a visitor enters an SSL-protected website, the SSL certificate automatically creates a secure, encrypted connection with their browser. SSL certificates enable websites to move from HTTP to HTTPS, which is more secure. An SSL certificate is a data file hosted in a website's server. SSL certificates make SSL/TLS encryption possible, and they contain the website's public key and the website's identity, along with related information.

We have installed an SSL certificate from SSL for Free (https://www.sslforfree.com/) which is powered by ZeroSSL (https://zerossl.com/). This is a zero cost SSL certificate, which needs to be renewed after every 90 days and is trusted by all major web browsers worldwide.

We are considering getting the SSL certificate from Microsoft Azure portal, however that involves additional cost and will depend on the availability of funds and sustainability options.

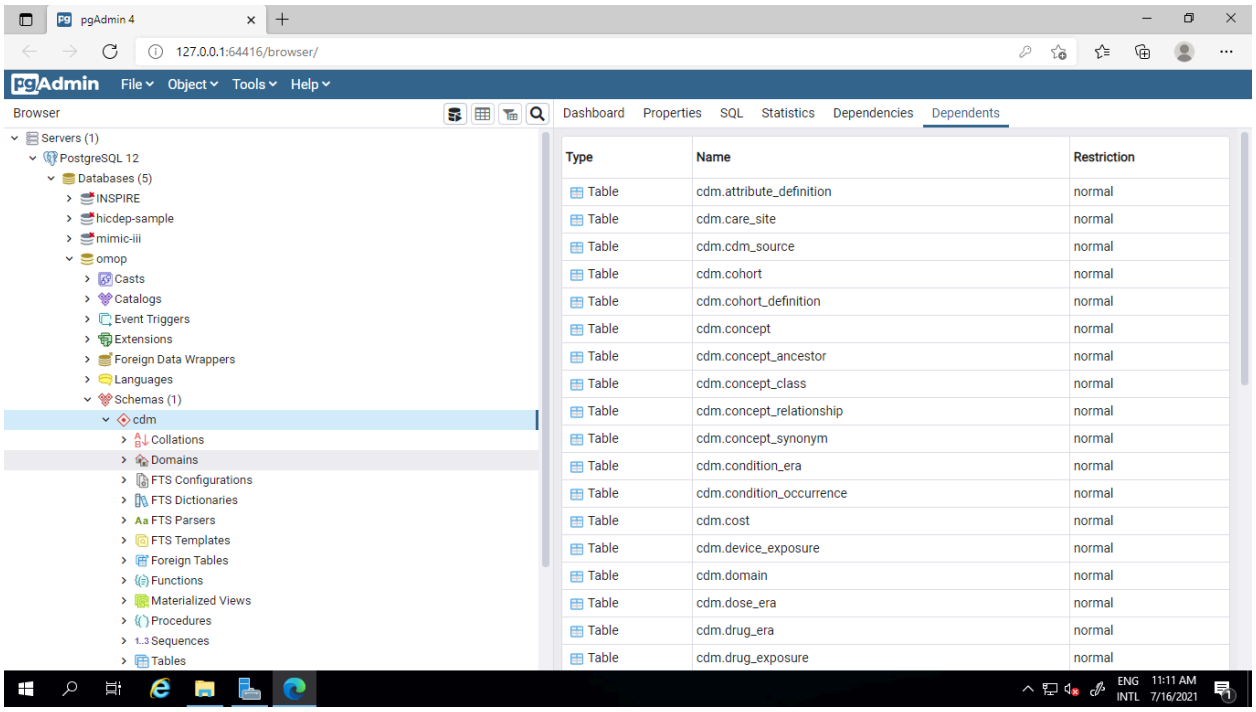
PostgreSQL Database with pgAdmin

PostgreSQL is a powerful, advanced open-source, object-relational database management system (RDBMS). It has a strong reputation for reliability, feature robustness, and performance. We had chosen to install PostgreSQL on our server because OHDSI tools strongly support this database. We installed PostgreSQL version: 12.4.

Along with PostgreSQL, we have installed the pgAdmin tool. It is the most popular and feature rich open-source administration and development platform for PostgreSQL.

A few sample databases were created for learning purposes. The OMOP vocabulary database was also populated here.

The following screenshot shows the pgAdmin home screen on the VM.



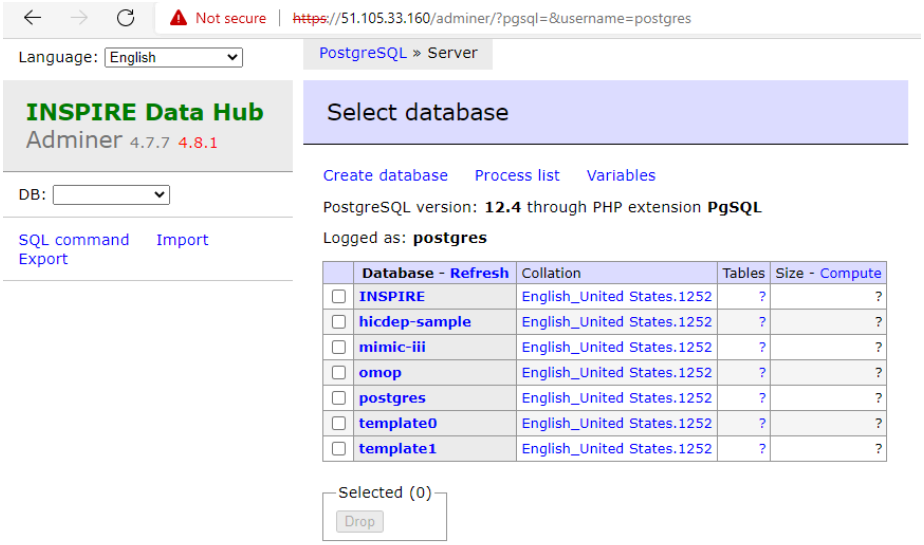
Adminer Database Management Tool

Adminer is a full-featured database management tool written in PHP. It is a web-based tool that can be accessed from the web browser and no additional installation is needed on the client computer. Adminer

can connect to many databases like MySQL, MariaDB, PostgreSQL, SQLite, MS SQL, Oracle, Elasticsearch, MongoDB and others via plugin.

We have installed Adminer to provide access to the database development and management team so that they can do it remotely using web browsers.

The following screenshot shows the Adminer home screen after login.



OMOP Vocabulary

The OMOP Standardized Vocabularies, or simply “the Vocabulary”, are a foundational part of the OHDSI research network, and an integral part of the CDM. It allows standardization of methods, definitions, and results by defining the content of the data. OHDSI requires harmonization not only to a standardized format, but also to a rigorous standard content.

The first step here was to download the OMOP CDM vocabulary from ATHENA website (https://athena.ohdsi.org/). To do that you must first register on the ATHENA website and then, after successful registration, you must log in and download the vocabularies in csv files.

We have created the OMOP database in PostgreSQL by running the following scripts from https://github.com/OHDSI/CommonDataModel/tree/master/PostgreSQL.

1. PostgreSQL script to create OMOP common data model results schema version 6.0

   OMOP CDM Results postgresql ddl.txt

2. PostgreSQL script to create foreign key, unique, and check constraints within the OMOP common data model, version 6.0

   OMOP CDM postgresql constraints.txt

3. PostgreSQL script to create OMOP common data model version 6.0

OMOP CDM postgresql ddl.txt

4. PostgreSQL script to create the required primary keys and indices within the OMOP common data model, version 6.0

   OMOP CDM postgresql pk indexes.txt

Executed the SQL statements from the scripts in the following order to create the OMOP database table structures and then populate it with the vocabularies.

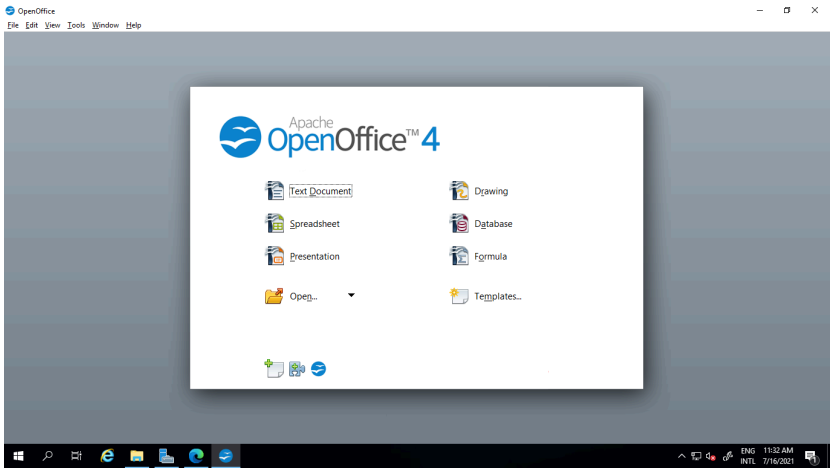
1. Created an empty schema named omop
2. Executed the script "*OMOP CDM postgresql ddl.txt"* to create the database tables and fields in the omop schema.
3. Loaded data into tables in the omop schema. To load data, the script *"OMOP CDM vocabulary load - PostgreSQL.sql"* was executed after making the necessary changes in csv file paths that were downloaded from ATHENA website (https://athena.ohdsi.org/). This file uses the COPY command to copy the data from the csv files to PostgreSQL tables.
4. Executed the script "*OMOP CDM postgresql pk indexes.txt*" to add the minimum set of indexes and primary keys.
5. Executed the script "*OMOP CDM postgresql constraints.txt"* to add the constraints (foreign keys).

Apache OpenOffice

Apache OpenOffice (http://www.openoffice.org/) is an open-source office productivity software suite. It contains a word processor (Writer), a spreadsheet (Calc), a presentation application (Impress), a drawing application (Draw), a formula editor (Math), and a database management application (Base).

We installed OpenOffice as part of our effort to remain with open-source products as much as possible for the INSPIRE data hub. Various data formats can be opened for work with OpenOffice tools like Calc and Base instead of MS Excel and MS Access and for any documentation work on the VM. Apache OpenOffice 4.1.7 was installed.

The following screenshot shows the home screen of OpenOffice in the VM.



OHDSI White Rabbit

White Rabbit is an open-source tool from the OHDSI stack. It is Java based and thus platform independent.

White Rabbit helps to prepare ETLs (Extraction, Transformation, Loading) of longitudinal healthcare databases into the OMOP Common Data Model (CDM). The source data can be in comma-separated text files or in a database (MySQL, SQL Server, ORACLE, PostgreSQL); the CDM will be in a database (MySQL, SQL Server, PostgreSQL).

The main function of WhiteRabbit is to perform a scan of the source data, providing detailed information on the tables, fields, and values that appear in a field. This scan will generate a report that can be used as a reference when designing the ETL, for instance when using the Rabbit-In-a-Hat tool.

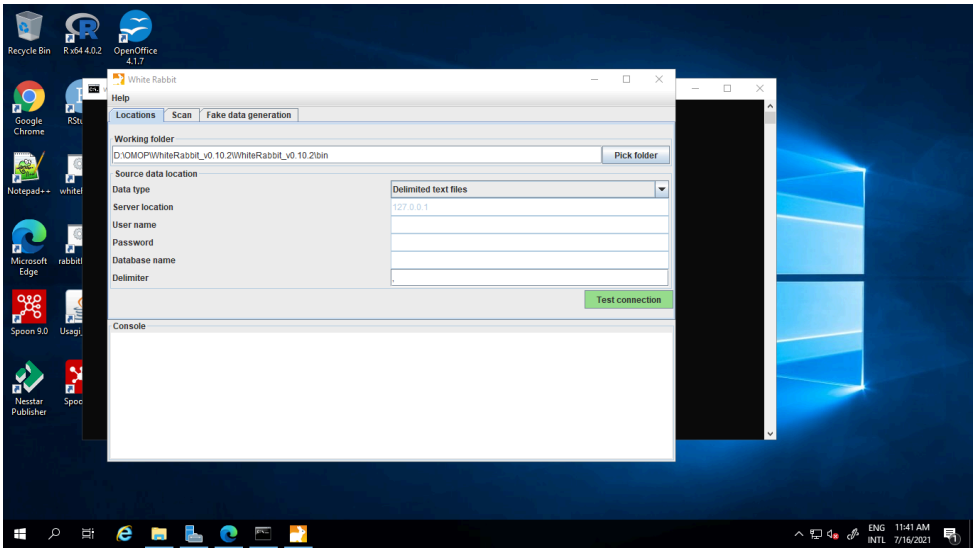
At INSPIRE, to map different source datasets to OMOP CDM, we use White Rabbit for initial profiling of the dataset and create the scan report for mapping purposes.

Process of installation:

http://ohdsi.github.io/WhiteRabbit/WhiteRabbit.html#installation_and_support

Download URL: https://github.com/OHDSI/WhiteRabbit/releases/tag/v0.10.2

The following screenshot shows White Rabbit running in the VM.



OHDSI Rabbit in a Hat

Rabbit in a Hat is an open-source tool from the OHDSI stack. It is Java based and thus platform independent.

Rabbit-In-a-Hat comes with White Rabbit and is designed to use the scanned documents generated in White Rabbit to display the source data information through a graphical user interface to allow a user to

connect source data structure to the OMOP CDM data structure. The function of Rabbit-In-a-Hat is to generate documentation for the ETL process, not generate code to create an ETL.

At INSPIRE, we use Rabbit in a Hat to generate the mapping documents which assists ETL developers to create the ETL processes for moving data from different sources to OMOP CDM.

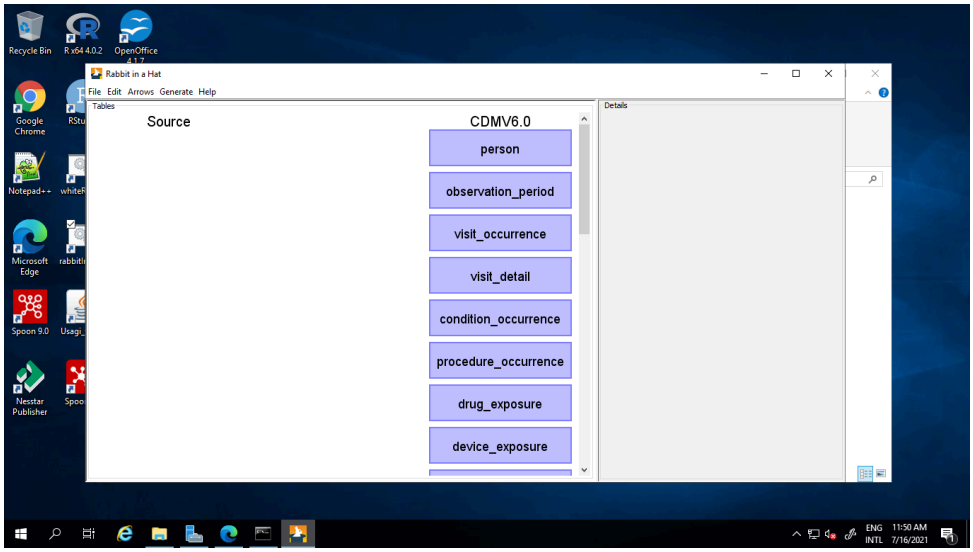
The Rabbit-in-a-Hat tool comes with White Rabbit software, and separate download is not needed.

Process of installation:

http://ohdsi.github.io/WhiteRabbit/WhiteRabbit.html#installation_and_support

Download URL: https://github.com/OHDSI/WhiteRabbit/releases/tag/v0.10.2

The following screenshot shows Rabbit in a Hat running in the VM.



OHDSI Usagi

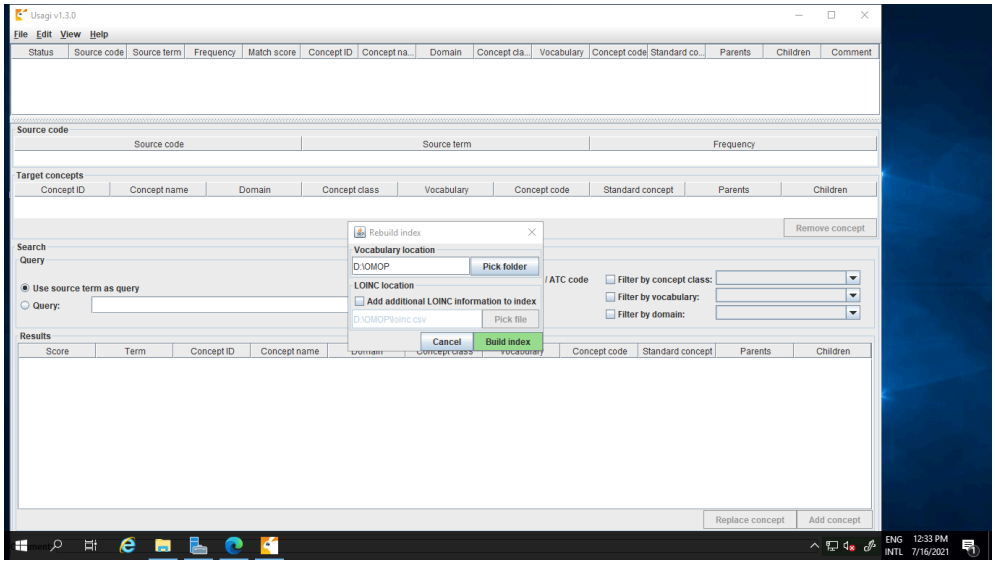
Usagi is an open-source tool from the OHDSI stack. It is Java based and thus platform independent. It too can run on Windows, Linux, and Mac.

Usagi is used to help in the process of mapping codes from a source system into the standard terminologies stored in the OMOP vocabulary.

At INSPIRE, we use Usagi for mapping some source codes to the OMOP vocabulary.

Process of installation: https://github.com/OHDSI/Usagi#getting-started

The following screenshot shows Usagi running in the VM.



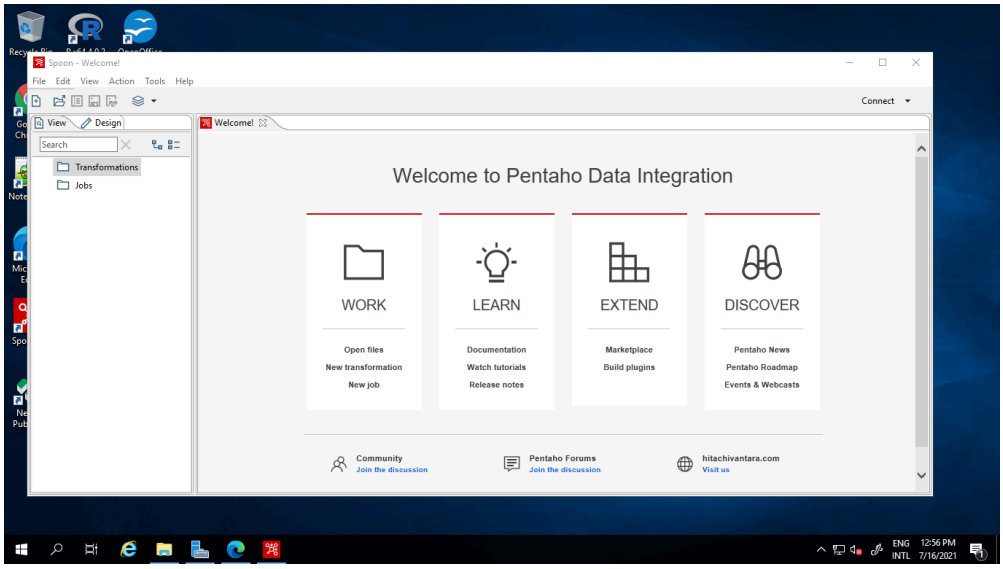
Pentaho Data Integration

Pentaho Data Integration (PDI), also known as Kettle, provides the Extract, Transform, and Load (ETL) capabilities that facilitate the process of capturing, cleansing, and storing data using a uniform and consistent format that is accessible and relevant to end users. A PDI client (also known as Spoon) is a desktop application that enables the user to build transformations as well as schedule and run jobs with a GUI interface.

We have installed PDI to create the ETL pipeline for implementation of data mapping that was documented using Rabbit in a Hat.

We have installed PDI 9.0 Community Edition. For more details on installation: https://wiki.pentaho.com/

The following screenshot shows a PDI spoon running in the VM.

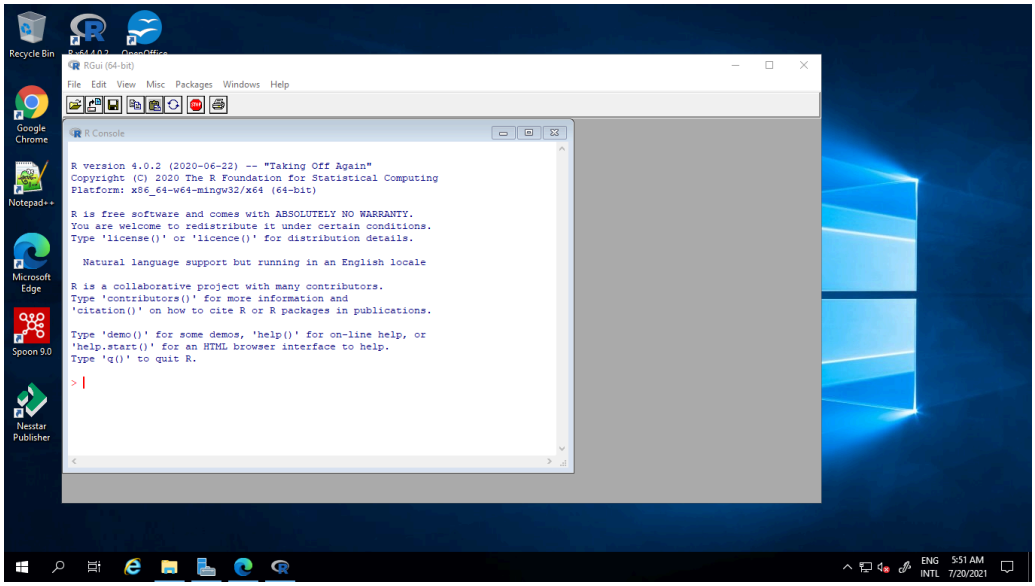


R

R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.

We have installed R to allow researchers and data scientists of INSPIRE EA to do data analysis using this free environment. Also, many OHDSI tools are based on the R programming language. In order to facilitate the use of those tools, it was necessary to have R installed on the server. We have installed R x64 4.0.2 for Windows.

The following screenshot shows R running in the VM.

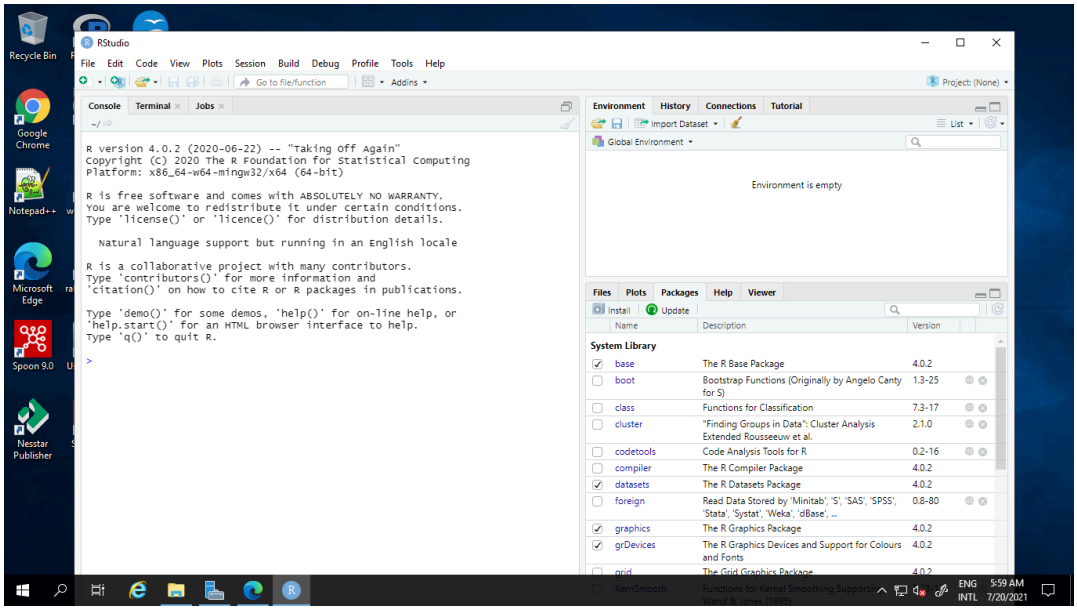


RStudio

RStudio is an Integrated Development Environment (IDE) for R programming language. RStudio is available in two formats: RStudio Desktop which is a desktop application and RStudio server which runs on a remote server and allows it to be used through a web browser.

We have installed RStudio 1.3.1073 for Windows desktop. This will facilitate researchers and data scientists of INSPIRE EA who log into the VM, using remote desktop connection, to use this IDE for R programming.

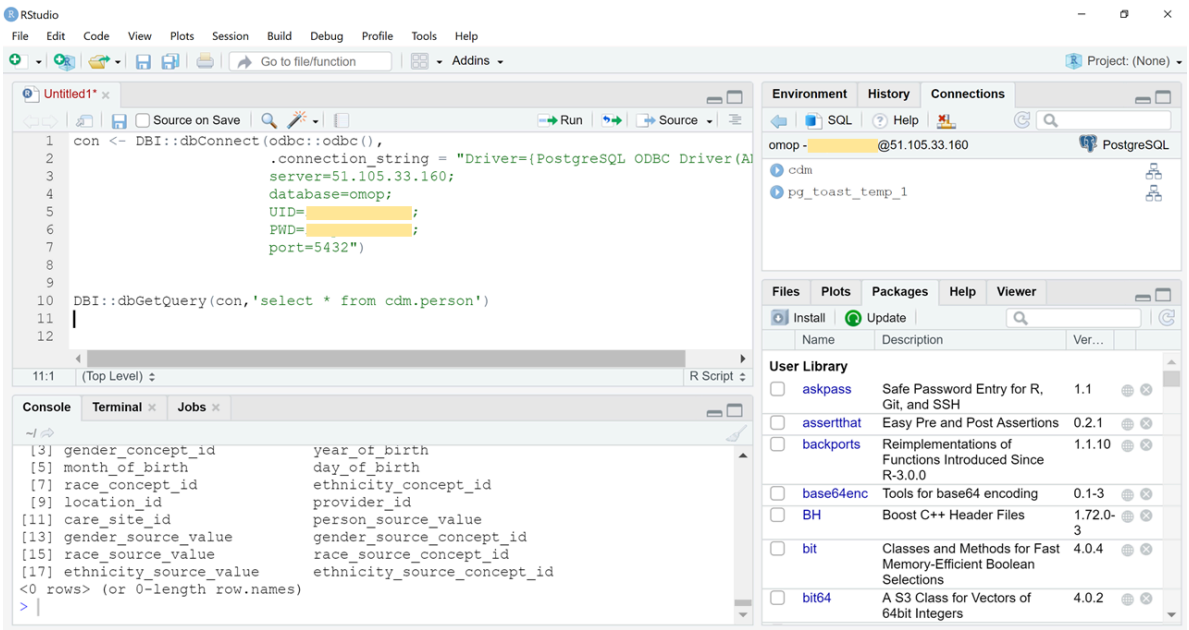
The following screenshot shows RStudio running in the VM.



Remote connection to PostgreSQL Database Using R

R users can remotely connect to our PostgreSQL server. The OHDSI tools that are installed on remote computers can also connect to the OMOP CDM on the VM and do the necessary off-ramping and data analytics work. INSPIRE EA team members are given PostgreSQL user credentials on the OMOP CDM datasets to carry out the tasks.

The following screenshot shows RStudio running in a remote computer and connecting to the PostgreSQL database on the VM.

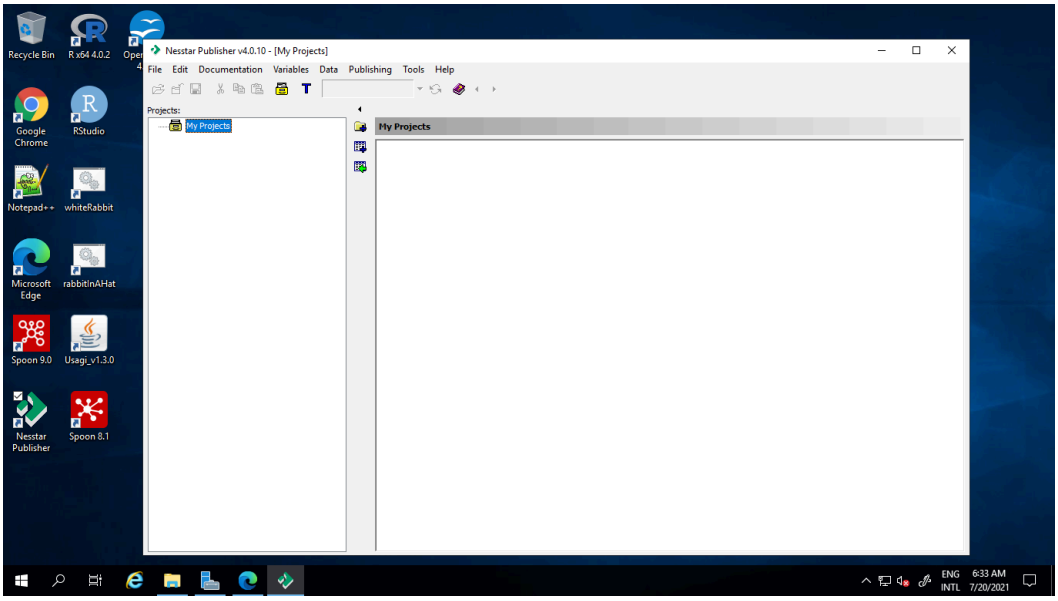


Nesstar Publisher

Nesstar Publisher (http://ihsn.org/software/ddi-metadata-editor) is a metadata editor from International Household Survey Network (ISHN - http://ihsn.org/). It is a feature rich editor for the preparation of metadata and data for publishing in an online catalog such as the ISHN developed National Data Archive (NADA - http://nada.ihsn.org/). The metadata produced by the Nesstar Publisher editor is compliant with the Data Documentation Initiative (DDI) 2 (https://ddialliance.org/) and the Dublin Core XML metadata standards. The application was developed by Nesstar at the Norwegian Social Science Data Archive (NSD) and is distributed as freeware.

We have installed Nesstar Publisher to facilitate users who log into the VM using remote desktop connection to prepare metadata of the source and target datasets.

The following screenshot shows Nesstar Publisher running in the VM.

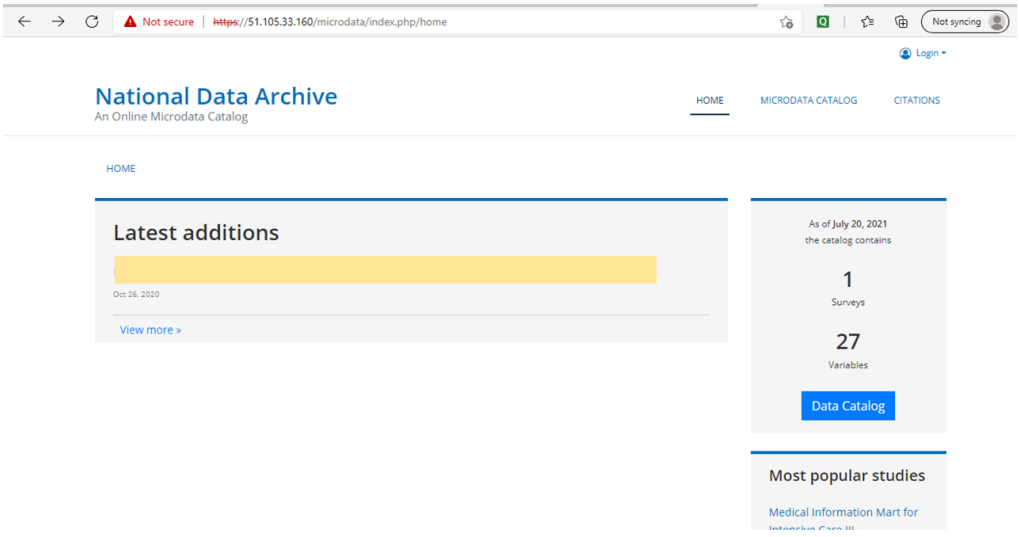


National Data Archive (Online Microdata Catalog)

It is an online microdata cataloging tool. It is an open-source software designed for researchers to browse, search, compare, apply for access, and download research data. We have installed the NADA catalog to facilitate the hosting of metadata produced by Nesstar Publisher and to share datasets.

We have installed NADA 5.0.4. This resource can be accessed remotely using any web browser with the following URL: https://51.105.33.160/microdata/index.php/home

The following screenshot shows the online data catalog accessed remotely using a web browser.



MySQL

MySQL is an open-source relational database management system (RDBMS) which is used to store and retrieve data efficiently. It is free and open-source, making it ideal for small to medium sized database applications. It is often used as back-end databases for web-based applications.

We have installed MySQL to facilitate the installation of the NADA online microdata cataloging tool. It can also be used to host any other dataset deemed fit for this purpose. The Adminer browser-based tool (described in an above section) is to be used to access this database remotely from our server. We have installed MySQL 5.7 on Windows.

Conclusion

This infrastructure was created on Microsoft Azure cloud service to enable the hosting of INSPIRE EA data hub. The data hub had two primary objectives. (a) To provide a platform to load data on the OMOP CDM from different input sources. This process is known as the data on-ramp to the INSPIRE EA data hub. (b) To access data from the OMOP CDM on INSPIRE EA's data hub, different host based or remote and web-based applications from the OHDSI stack can be used. Alternatively, other applications can also be used for accessing the data. This process is known as the off-ramp from INSPIRE EA data hub.

This infrastructure provided a platform for developing an environment for deploying the OHDSI stack as a way of managing population and disease data from INSPIRE EA member sites. This implementation under Phase I of the project is not a complete solution by itself, but is a pilot initiative in achieving the objectives of the project.

INSPIRE Data Hub: High-Level Architecture and Data and Metadata Flows

Version 0.1

Version 1.0

Version	Date	Author	Reviewer	Description
1.0	2021-07-14	Tathagata Bhattacharjee		This version outlines the processes involved in setting up the IT infrastructure on Cloud for INSPIRE EA.

Contents

I. Overview

The INSPIRE Data Hub is a FAIR data resource containing longitudinal population health data from Health and Demographic Surveillance System (HDSS) sites in southern and eastern Africa. It is designed with the idea that population health data can be usefully combined with data from other sources, notably routine healthcare data from clinics. It is designed to be both scalable and extensible, allowing for additional data in new areas to be introduced without requiring a new hub infrastructure. To facilitate this goal, the hub is based on international standards for describing both population-based health data and clinical data.

New types of data (e.g., results of genomic sequencing) will be integrated with the platform as this becomes possible, and the general scope will be broadened. The central data model for the hub is based on a standard which is currently being extended to support such integrations by others – the OMOP Common Data Model (CDM) – and the hub will incorporate this new functionality as it becomes established. To this end, INSPIRE is working toward the establishment of an African chapter of OHDSI, the group which coordinates such developments.

The long-term goal is to provide a robust pan-African platform for data integration. The hub is designed to coordinate with other African initiatives such as the national network service providers (NRENs) and the African Open Science Platform (AOSP). The INSPIRE Data Hub is being developed by a network of HDSS sites and interested organizations which can provide the needed governance for the platform. This document does not describe the network or organizational aspects of INSPIRE but focuses on the technical architecture and related developments.

In the longer term, it is hoped that the INSPIRE Data Hub can become part of an international network for FAIR data sharing. Given the scientific challenges of dealing with phenomenon such as the recent global pandemic, the need for such data sharing is manifest. The use of internationally recognized standards and vocabularies is a key technical component to ensuring that this remains a possibility.

In the short term, INSPIRE must first establish the basis for such a data-sharing hub service, and this document provides a report on the initial developments in this area. It will introduce the various system

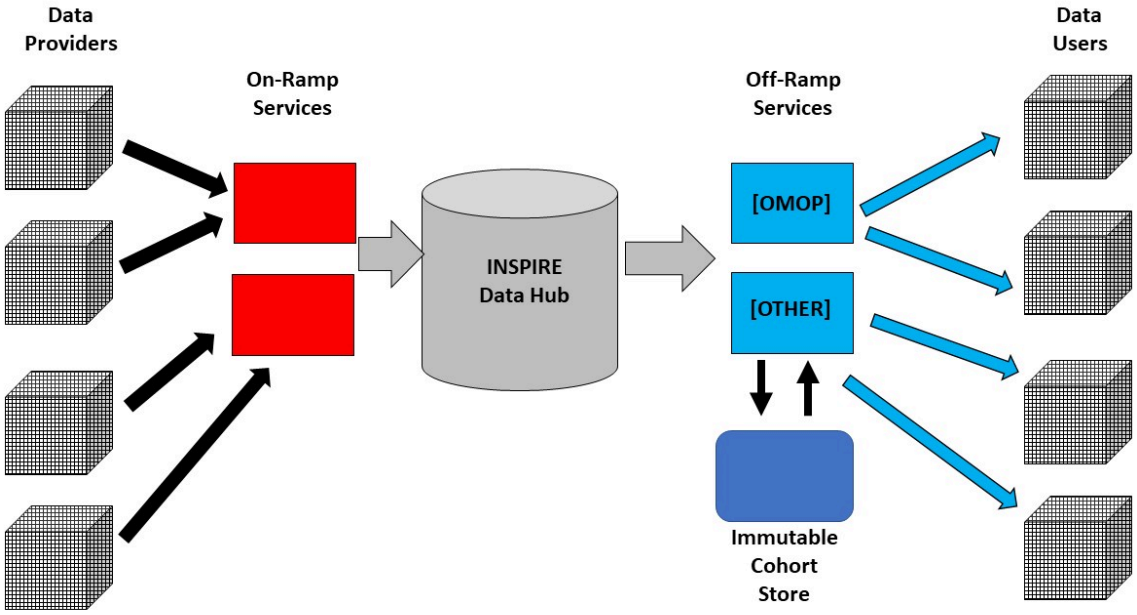
components, and examine how the data is modelled, brought into the system, and then disseminated. Support for data integration is a key, involving the mapping of concepts and vocabularies, but further providing the context by which the origins of specific data can be understood for use in analysis. Issues such as access and disclosure control are also addressed.

This document reflects the system as designed and prototypes in Phase I of the INSPIRE project. Not all of the designed functionality has been prototyped, but the deployment of the core data model and some of the associated services has been performed and has informed the overall approach. It should be noted that this work builds on earlier developments in the HDSS space, including those of the INDEPTH and ALPHA Networks.

II. SYSTEM COMPONENTS

This section provides an overview of the major components of the system, and their basic roles and functions. The hub itself is a "platform as a service" (PaaS) which is cloud-based, helping to address issues of scalability, facilitating data back-ups, and other basic administrative functions. The initial prototype used a commercial cloud service provider, and an assessment was conducted to identify which of these was most suitable for an African data resource. Findings are recorded [LINK TO TATHAGATA'S DOCUMENT HERE]. It is expected that the NRENs will provide a long-term solution for this aspect of the system, but that has yet to be determined, and is beyond the scope of the initial system prototyping and development.

The diagram below shows the major system components at a conceptual level. These will be characterized, and their functions described in more detail below.



Conceptual Components of the INSPIRE Architecture

Data Providers

Data providers are those institutions which have an agreement to supply data to the hub. This agreement will specify the on-ramp (the technical specifications) for how that data is supplied, including the format, protocols, validation criteria, and needed documentation.

On-Ramp Services

On-ramp services are those services couched in terms of community-specific standards for the exchange of data and metadata/documentation, and any other information required to facilitate the transmission of data to the hub. While the inputs will exist according to the standards in use for a specific community or according to a specific set of standards, the outputs from the service will be mapped to the internal hub data model (its implementation of the OMOP CDM – see below). Thus, any given on-ramp is essentially an "extraction, transformation, and load" (ETL) process in traditional IT terminology.

INSPIRE Data Hub

The INSPIRE Data Hub is a database which implements the OMOP CDM, with some specializations to support the inclusion of population data and the flow of documentation/metadata needed (see below). It is designed to be a compliant implementation of the OMOP specification, such that the full range of tools from OHDSI will work as they would on other OMOP implementations. The hub is a cloud-based implementation of an open-source relational database (Postgres), and builds on top of that system's administrative capabilities and functionality.

Conceptually, the Data Hub is not the point where the data is managed but acts as a service for providing data which is owned and managed elsewhere (by the data providers) in a form which allows for it to be easily used and integrated with other data. In this sense the Hub is a platform-as-a-service (PaaS) rather than a data management or dissemination application. It serves as the platform on which applications for data dissemination and analysis can be built, and many such applications – those which natively work with the OMOP CDM – already exist.

Despite the fact that it is not a data management application as such, it does need to perform many of the functions typically found in those systems: versioning, identification, access control, etc. The requirements for this functionality are a consequence of its function as a service platform, however, and not as a typical data management tool.

Off-Ramp Services

Off-ramps provide services for accessing the data and metadata held in the hub. Any given off-ramp service operates according to the standards (open or proprietary) of an intended target audience of users. Because the OMOP CDM is in common use for clinical research, the tools which support that standard represent a large user community. Other user communities are more used to other tools for analyzing data, with Stata being a common choice.

Thus, off-ramps may include processing to the information found in the Hub, in order to provide it in a useful fashion. This includes not just the data, but also the needed metadata and documentation. Off-ramps may rely on commonly used "delivery" standards such as the Data Documentation Initiative (DDI), which supports a range of tools for producing analysis packages for a range of statistical packages

in common use in social research (Stata, R, SPSS; SAS), as well as the needed "codebook" documentation.

Off-ramp services are provided to known users, whose access to the data holding of the Hub are controlled and subject to appropriate licensing and restrictions.

Immutable Cohort Store

Clinical research uses the idea of "cohorts" – definitions of which records are to be included for analysis based on values for time and other data points within each record. The Hub implements this concept as it exists within OMOP CDM. Such cohort definitions can be created such that the data set is updated over time, and automatically reflects changes in any values contained in the data set as a result of corrections. Some types of research (and thus off-ramps serving them) operate on the basis of static analysis data sets, which are also important for purposes of disclosure risk control, citation, and reproducibility of findings. The Immutable Cohort Data Store supports this requirement by providing an ability to recreate any given data set drawn from the Hub by persisting information about the state of the Hub data when the cohort was applied. Such "immutable cohorts" can thus support the many functions which rely on the existence of static, unchanging versions of data sets. While not needed for all off-ramps (OMOP-based applications provide their own solutions to these problems) the ability to support different users' needs requires this functionality.

Data Users

Data users will be any individual or application which uses the data and metadata found in the Hub. Users will be known to the system and will be granted appropriate access based on their role and accreditation. Any given user may access one or more of the off-ramps, based on what protocols and standard they support. Data users may be both individuals and, potentially, applications with their own community of users which are trusted to act responsibly. It is also possible that there will be another class of users, more concerned with the metadata than with the data – one can imagine that search providers might wish to index the holdings of the Hub but have no requirement to access the data itself. Specific off-ramps could support these types of use, as well as the more typical use of the Hub as a source of data for analysis.

III. INTAKE OF DATA: ON-RAMPS

The data flows through the Hub are simple on a conceptual level: data is submitted through an on-ramp, an ETL process is performed, the results validated, and the data is then ready for access by any interested user through the off-ramp services and applications which use them. This simple flow requires that some challenging tasks be performed in order for it to function.

The biggest hurdle is the rendering of the data into a harmonized form which allows it to be freely reused. Key to this is the data model on which the Hub itself is based: the OMOP Common Data Model. This model essentially associates every observation with a Concept, generally taken from a standard vocabulary, of which there is a centrally maintained list in an OHDSI registry called Athena (https://athena.ohdsi.org/). The typical vocabularies are well-established ones such as SNOMED, LOINC, and so on. It is also possible for non-standard vocabularies to be used (they are entered into Athena so they are available). There is further a type of vocabulary known as a "classification" vocabulary which

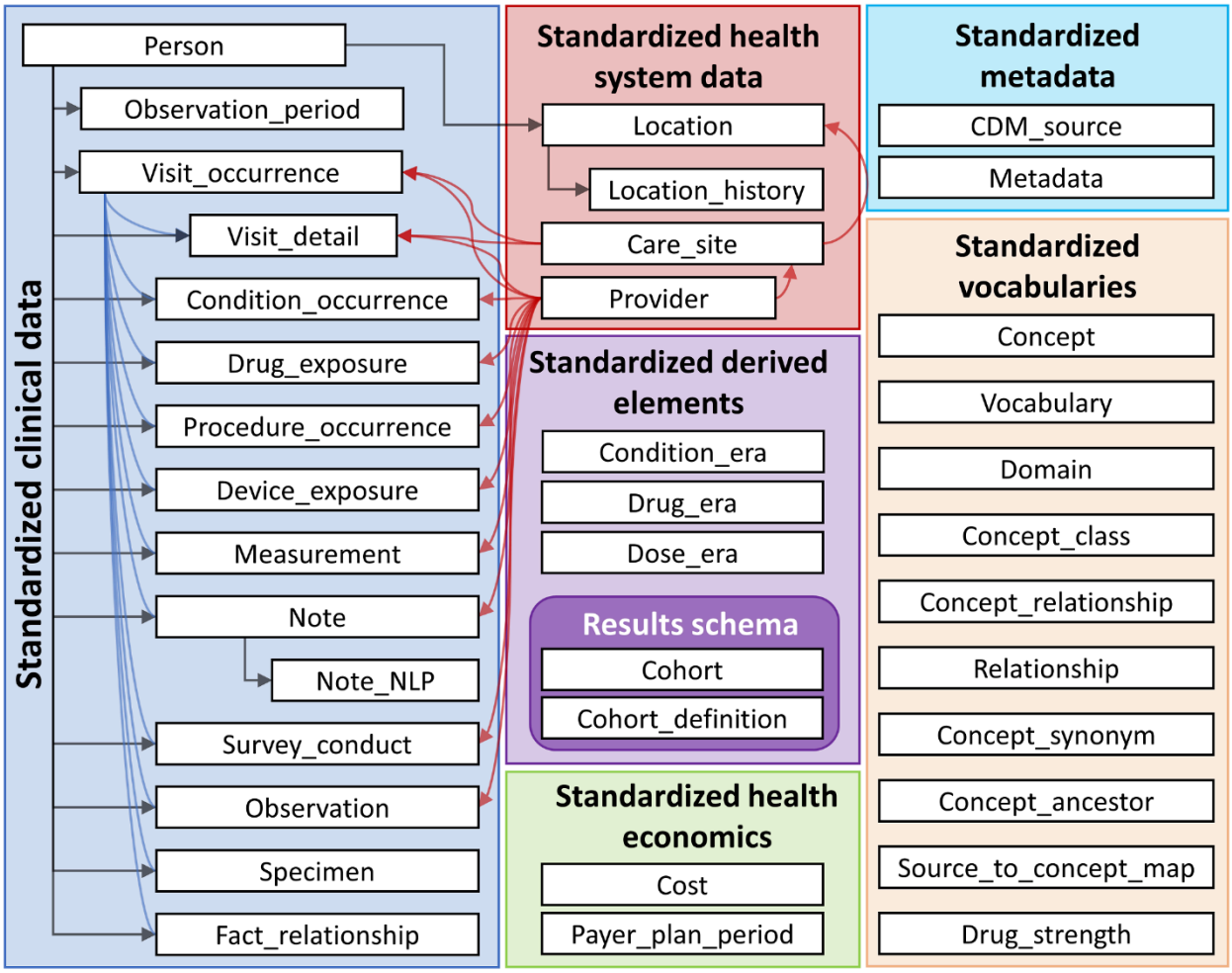
serves to group related concepts used directly by observations, allowing hierarchical vocabularies to be described.

OMOP is at base a relational type of model, which is familiar to many users. There are a number of standard tables, with agreed columns, so that data can be mapped into a standard structure, with the observations represented by the standard concepts given above. The values for codes are numeric strings assigned by the OMOP CDM; rather than the native source codes, although these may be stored for reference. It provides a flexible and approachable way for clinical data to be shared effectively, based on many of the common standards in use today.

Many of the standard tables and vocabularies within the OMOP CDM are focused on clinical research, however, which was not in all cases sufficient for the mapping of population data. The resulting work identified a way to describe population data, using HIV data from members of the ALPHA Network HDSS sites as a test case. (In essence, the shared ALPHA data specifications already in use became the standard format for data and metadata used by that particular on-ramp.)

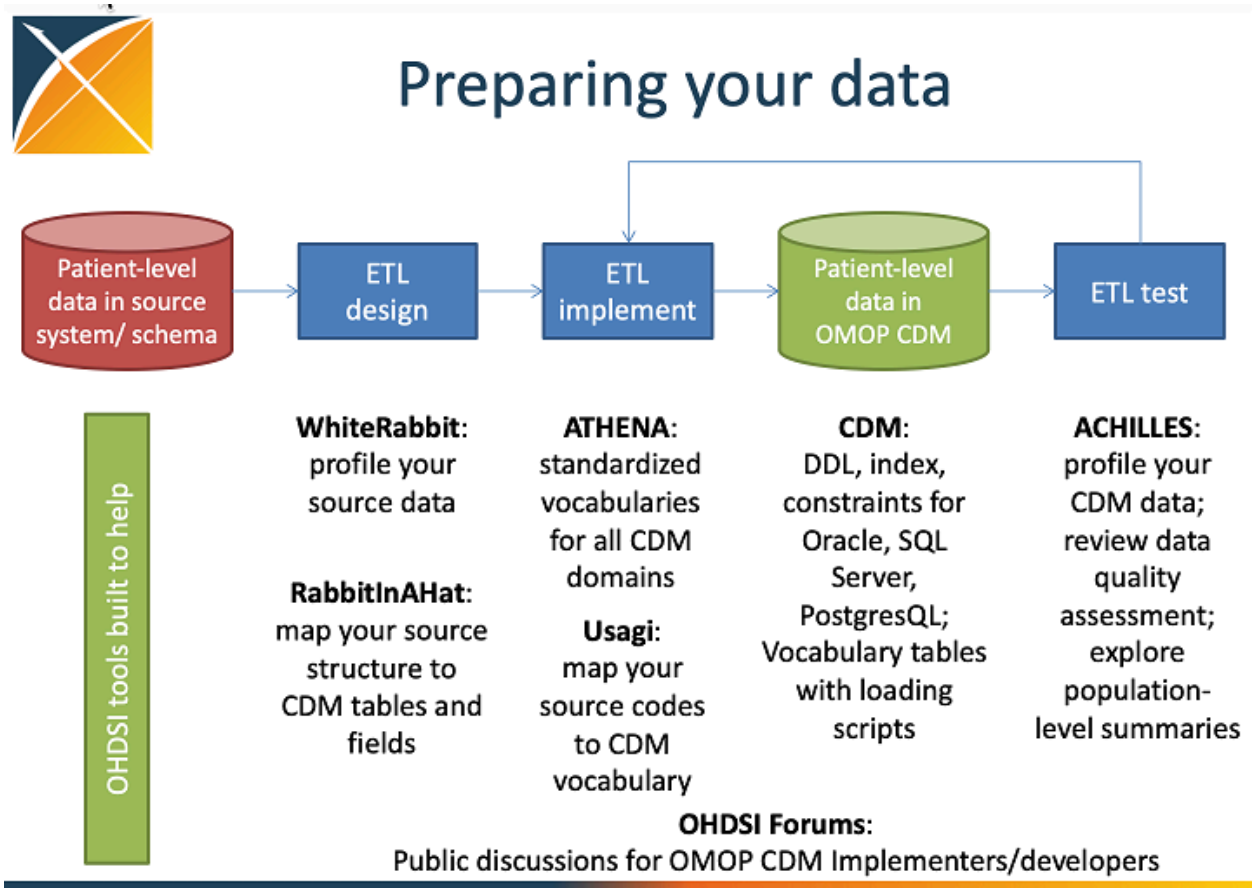
Ultimately, this resulted in the creation of non-standard vocabularies within the OMOP CDM sufficient to use the basic CDM structure, but with the needed additional concepts in those cases where they did not already exist. To facilitate this process, the existing OHDSI mapping tools White Rabbit and Rabbit in a Hat were employed. Discussions were initiated with OHDSI to understand how to establish the new vocabularies required for population data to become available within the Athena registry.

The *Book of OHDSI* is a publication which describes the OMOP model, and it provides the following high-level diagram which gives a flavor of how data is organized according to the CDM (from https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html):



Each of the white boxes is a standard table with a series of columns, the value for each being taken from a recognized concept based on a (typically standardized) vocabulary, themselves described within the data model as shown.

The way in which ODHSI applications built on OMOP may be employed to achieve this task is shown in the diagram below.



This general approach was used in mapping the population data for the INSPIRE prototype into the OMOP CDM, A detailed discussion of the mapping process to describe population data with this model, and the results of that effort, are described in [CITE JAY AND TATHAGATA's DOCUMENT HERE].

While the INSPIRE prototype used sample HIV data formatted according to the ALPHA Network specifications as the primary test input for this process, it is on which can be applied to a broad range of data sources. The intention of this design is to support as many community standards (such as the ALPHA Network specifications) as is reasonable, to bring data into the INSPIRE Hub. Each on-ramp thus becomes a community gateway for data to be provided. It is not the intention that every possible data source will be mapped in this way, but that those formats which are already agreed among a set of data providers be supported.

Because such community specifications will already be understood, the problem of data acquisition can be reduced to a significant extent. Mapping work such as that described would be conducted in collaboration with the community experts for any given specification to be supported in an on-ramp.

III. METADATA: PROVIDERS AND DOCUMENTATION FLOWS

A. Overview

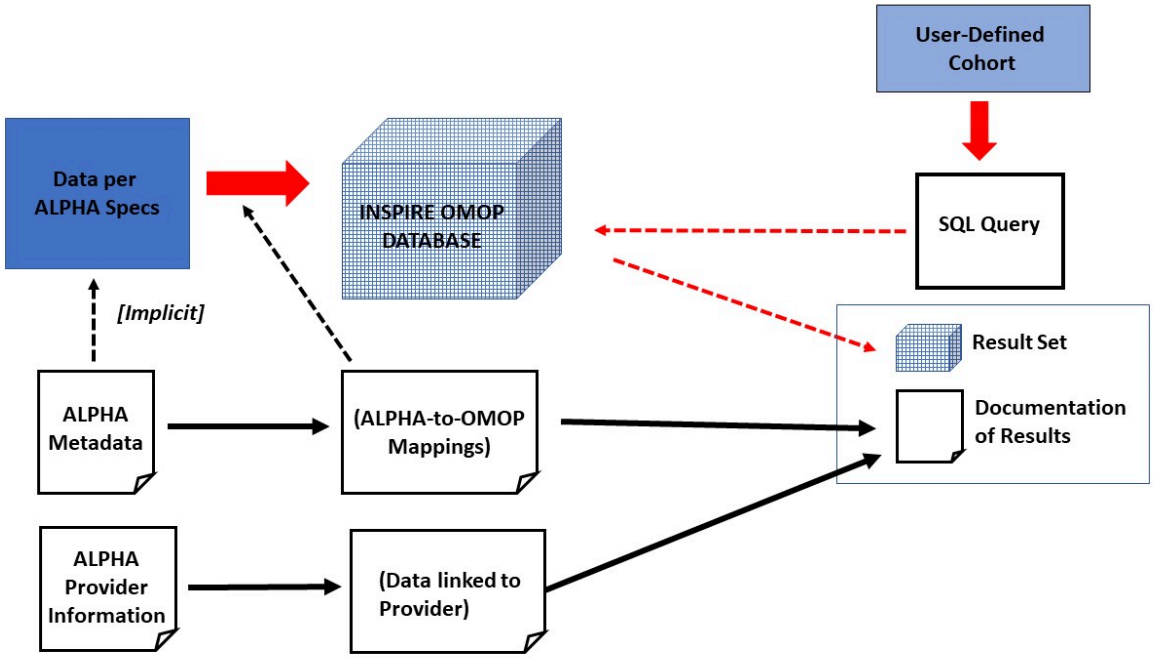
This section outlines the flow of documentation through the INSPIRE Hub. This is presented first at a conceptual level, and then in terms of more detailed considerations: the metadata inherent in the ETLs targeting the OMOP CDM model from input formats, and the additional information which will need to be captured at a higher level for documenting the data when used. Mechanisms for exposing this information to users through the different "off-ramps" is described in the following section.

The ODHSI research paradigm is based on several assumptions about the data being made available, which are familiar to users. Each instance of the OMOPCDM is assumed to be populated from an institutional source, and knowledge of that source is assumed on the part of the users. INSPIRE – because it combines data from many sources – does not fit neatly into this paradigm. Within this data ecosystem, information about the data at a granular level is well-described, notably the Athena registry of vocabularies and the standard tables of the CDM itself provide a wealth of definitional information. In addition, there are many documented conventions of use to help those working with the data to understand it.

For this class of users, it is possible to provide additional background documentation, but this function is not a required aspect of the system as designed and is not incorporated explicitly in the OMOP CDM. For users operating outside this environment, more context and information must be provided and expressed in a fashion similar to what would be found in typical non-OHDSI research settings. This section will look at what information could be captured at which stage of the process, to support the delivery of data effectively to different audiences accompanied by sufficient documentation and metadata to permit effective use. The delivery of metadata through off-ramps to different target audiences will be considered in the following section.

B. High-Level/Conceptual View

The diagram below shows an example of the conceptual flow of metadata used for the purposes of documentation through the INSPIRE Hub:



This diagram serves to provide a frame within which we can understand the overall flows of metadata.

The red arrows show the flow of data (including a significant proportion of needed metadata):

- Data conforming to an ALPHA specification is subject to an ETL rendering it as valid data formatted according to the Hub's OMOP CDM model. Each cell in the input data is mapped to a corresponding column in the standard OMOP CDM tables, and the coding in ALPHA is transformed to the corresponding concept taken from a standard vocabulary.

- A user has defined a Cohort (in OMOP terms) which can be expressed (or directly written) as a SQL query. When executed. This query produces a result set including the needed data (and at least part of the metadata) needed by the user.

The black arrows show the documentation flow.

- Implicit in the ALPHA specifications are a set of agreed definitions, which have been documented separately. This documentation exists external to the ETL mechanism used to submit the data itself but is available in other formats (e.g., can be expressed in a DDI Codebook XML format).

- The mappings between the ALPHA data and the Hub OMOP database structures are themselves reliant on the metadata which inform those mappings. This includes not only definition of relevant concepts and their representations, but also some aspects of methodology (e.g., how an episode is defined.) This information exists explicitly as documentation of the mappings, as well as in machine-accessible systems (within the ETL platform, the OMOP CDM tables and the *Athena* repository).

- Higher-level information about the institution and the data-collection efforts, along with additional general information regarding licensing and conditions of use, funding, methodology, and so on can be captured in a documentary form. This provider information is relatively static and should accompany the submission of data to the INSPIRE Hub. Each of the data points from that source would then be linked to the provider and the effort conducted by that provider to produce the data (the data production stream).

- When the data is queried, using a cohort (defined by the user in an OHDSI tool such as ATLAS, or pre-defined and stored for use in the INSPIRE Hub) expressed as an SQL query, the associated documentary data-level metadata (names, definitions, and sources for standard concepts, the definition of fields in the standard OMOP CDM tables) can be extracted and provided (if this function is not already being performed by the OHDSI tools). Additionally, and relevant higher-level information about the provider and their data can be assembled, based on the contents of the result set, and assembled and formatted for delivery.

- In order to understand how these processes can be implemented, we will need to consider the requirements of the intended audiences – this is addressed below.

C. Collection and Storage of Provider-Level Information

The conceptual flow above recognizes that documentary information regarding the provider of data must accompany the data itself, to be available to inform users of that data. This section looks more closely at the specific content of that data and considers how it might be stored and accessed in relation to the data points stored within the INSPIRE OMOP CDM instance.

The useful set of provider information, and information about the efforts to collect and process the data before submission to the INSPIRE Hub, can vary. A good general guide to the documentation and metadata useful from this perspective can be found in the recommendations made by the International Household Survey Network (IHSN) in their “Quick Reference Guide for Data Archivists” (https://guide-for-data-archivists.readthedocs.io/en/latest/). Of special interest is the section of that guide “7.2 Good practices for completing the Study Description.”

Each data point coming from any provider could have an entry within the OMOP CDM (using the METADATA table as described in the section below on off-ramps) linking it to the provider and the specific project or programme which produced it. For each producer, some or all of the information described in the IHSN guide above would be provided, stored in a static file format on the server which hosts the INSPIRE OMOP CDM instance. Such a format should be machine-processible (the DDI Codebook format would be an obvious candidate for this, bit other options are possible).

It should be noted that not all of the information recommended by the IHSN in their guide is relevant for the data contained in the INSPIRE Hub. Those recommendation assume that a single static data set is being described. Some of the information fields are not appropriate for use when the contents of the “data set” being described are produced as the result of a dynamic query (such as the definition and execution of an OMOP CDM cohort). Further, not all data providers will possess or desire to document all

of the information described. The IHSN recommendations are a general guide from which required and optional fields can be selected.

When a result set has been produced as the result of an OMOP CDM cohort execution, all of the relevant provider information can be assembled into documentary form for delivery to the end user. Because the data in any given cohort might come from multiple providers, each set of provider information could be assembled into a single presentation. The resulting document would serve to provide background for the specific data points but would not necessarily link each data point to its provider (although this could be done if desired). It is expected that descriptions of the location and coverage of the data collection efforts, given in general terms, will typically be enough to allow a user to understand which provider was the source of which data.

There are several sources of tools for visualizing DDI Codebook XML documentation in HTML and other formats. The IHSN provides these free of charge, as do several other organizations which use the DDI standards (see https://ddialliance.org/tool/ddi-xslt for an example of another such tool). None of these tools will provide the needed view of a list of data providers (they operate on the assumption that any given data set will have a single provider) but they offer a good starting point: a list of providers generated by the INSPIRE platform could link to the descriptions of each one creating using tools such as these, for example.

V. DISSEMINATION AND USE OF DATA: OFF-RAMPS

The prototype work focused on the on-ramps, as this presented the biggest barrier to effective data sharing from the Hub. Off-ramps for those applications which already understand the OMOP CDM are immediate candidates for deployment on the Hub, as it is a conforming OMOP application. The OMOP CDM is also well-supported by an R library designed specifically for working with OMOP data.

Other off-ramps were considered, and it was found that standards in common use for population research could be supported from the OMOP CDM. Notably, the Data Documentation Initiative (DDI) standards – often used to describe tabular data sets within the public health and social science domains – can be populated from the OMOP CDM as implemented in the INSPIRE Hub for describing data structures at a detailed level. This standard itself has good support from tools which allow the data to be formatted within many common analysis packages (e.g., R, Stata, SPSS, SAS).

One challenge for users is an understanding of the data model inherent in the tools. The OMOP CDM is a recognized model but is not necessarily familiar to all potential users of the data. For those who understand this model, OHDSI provides several tools of interest which can be used in analyzing and performing quality checks on the data, and these developments are on-going, spreading into new areas such as GIS displays.

The most common of these tools include:

ATLAS: This tool allows for users to define cohorts (the set of observations to be analyzed) in a point-and-click fashion and supports the definition and execution of some analysis functions.

ACHILLES: With this package, built in R, it is possible to do visualizations of data, and to produce summary reports to understand the data held across different sources.

OHDSI WebAPI: This is a set of web services which can be used for specific application development (ATLAS is implemented on top of this API).

HADES: This is a set of R libraries providing access to data held in OMOP CDM instances.

DATA QUALITY DASHBOARD: This is a tool which allows for reporting on the data held in an OMOP CDM instance according to a harmonized data quality terminology. It provides an assessment of the quality across data sources, rather than within a single selected set of data intended for analysis (for which other of the OHDSI tools are more appropriate).

These tools, along with *ATHENA* for accessing vocabularies and concepts, and the mapping tools mentioned above, are described on the OHDSI site (https://www.ohdsi.org/software-tools/).

For users who are familiar with the OMOP CDM and are comfortable with the OHDSI suite, these tools can all potentially be applied directly to the INSPIRE Hub. It should be noted that they require access to the data, and as such are subject to access control (see below). The use of these tools is supported by a basic set of training videos and materials provided by OHDSI, although these do not address the specific use of the OMOP CDM to address population data.

Some users will wish to use other tools to analyze the INSPIRE data, however, depending on their training and background. Within the research community which the INSPIRE Hub is expected to initially serve, STATA is a popular tool, and support for such users is understood as a requirement.

The OMOP CDM provides a solid framework from which to provide other formats. It is concept-rich, and has a clear, standards-based structure. The existing support for R also provides a strong indication that similar analysis packages can be supported. In the section below, we will look at how cohort definitions in the OHDSI framework can be used to formulate the metadata needed to prepare OMOP CDM data for analysis in such tools, as this primarily focuses on the structuring of the metadata rather than the data itself.

The bigger challenges encountered in the use of other analysis paradigms were not at the level of granular data description, but in the way that higher-level documentation about process, provenance, methodology, and so on were modelled. Within the OHDSI framework, a set of assumptions about data sources and use are in operation which provide sufficient information to researchers. The heart of such end-user documentation is in the *ATHENA* application, which provides a centralized place where all the definitions of concepts and their standard sources can be seen.

Users working outside this framework may want other forms of documentation, however, and these can also be supported through the use of existing structures within the OMOP CDM. The section below describes this approach.

VI. METADATA REQUIREMENTS AND FLOWS

A. Background

INSPIRE combines data from different types of sources, each of which has often traditionally been managed and used in different ways by different researchers. In order to understand the needs for

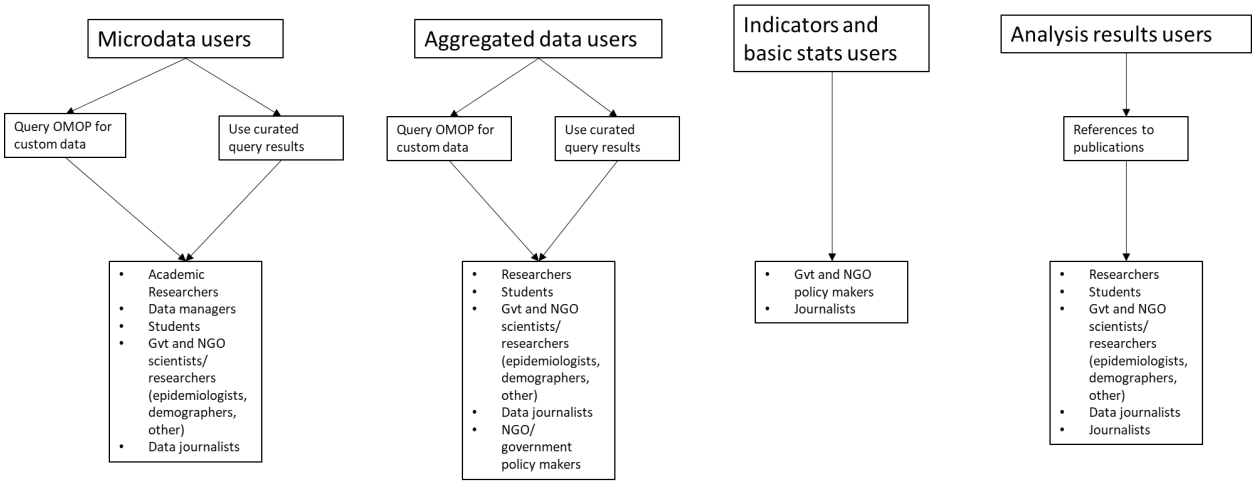
metadata and documentation across the breadth of the end-user community, it is useful to consider in broad terms the potential audience for this research data and the products of their work.

This section describes the analysis of the overall user community on which the INSPIRE work is based. While Phase I of the INSPIRE project focuses narrowly on only a portion of the overall audience, it is possible (and indeed, likely) that future work will broaden to support additional parts. The focus of Phase I is thus defined as targeting researchers interesting in accessing and analyzing the microdata available through the INSPIRE Hub for our current purposes. Further audiences can be understood within this general frame as the work carries forward in future.

This section first provides an overview of the general potential user community, and then specifically addresses the envisioned approach for serving specifically those researchers using the INSPIRE Hub as a source of microdata for analysis.

B. Potential Users

The diagram below shows the broad categorization of user types considered in this project:



The main categories are:

Microdata Users: "Microdata" is that data directly describing measurements, observations, responses (etc.) which are recorded at the level of individuals. The OMOP CDM holds data at this level in its PERSON tables. This is the level of detail of the greatest interest to researchers, as they will use data at this level to answer their specific research questions at a sufficient level of detail. Microdata forms the basis of episodes, visits to clinics, responses to surveys, and other events of interest to researchers.

Aggregated Data Users: Microdata can be tabulated and otherwise aggregated so that it is held at the level of populations and sub-populations (i.e., a count of all of the people within a given region with a specific diagnosis, etc.). This data is of interest to researchers, but usually as a reference for understanding the microdata which forms the primary material for their analysis. Other groups of users find this type of data to be more meaningful: policy makers, students, journalists, etc. The degree of sophistication needed to understand aggregate data is typically less than that needed to work with a

body of microdata – the individual performing the tabulation is assumed to understand the nuances of the input microdata on which the aggregation is based.

Indicators and Basic Statistics Users: This category builds further on the Aggregate Data one. Indicators are specific aggregate measurements which are fully defined and documented and tracked across time. Individual indicators often appear in various types of publications aimed at a generalist audience and are especially useful for comparing different phenomenon at a national, regional, and international level where the microdata may be collected and processed in a variety of ways. There are some other types of basic statistics which can also be useful in this way (i.e, non-aggregated summary statistics of various types).

Analysis Results Users: Most meaningful to general users, and critical as the outputs of research, are the findings and conclusions which the data support. While he INSPIRE Hub is not seen as a source for research papers or news articles, such publications often need to cite that data on which they were based. Given the scope of the INSPIURE Hub, it can be assumed that supporting such citations is a requisite function. Further, any discussion or disputation of such findings may require that an exact picture of the data used can be assembled after the fact, placing an archival requirement on the Hub. It is also common for such sources of combined data to advertise their quality by listing research papers driven by the resources they contain, as this is of benefit to the researchers themselves, and to the data platform as a resource.

Each of these categories is given more detail in the diagram above. In the case where data are supporting research directly (for microdata and aggregate data) the user communities are broken into two groups: those who are comfortable working within an OHDSI paradigm, which requires a working knowledge of the way in which the OMOP CDM describes data; and those who are working in another paradigm (e.g., using a statistical analysis tool such as STATA for their work). Both of these types of researchers are expected to be primary audiences for the INSPIRE Hub, and both are part of the focus for the design and architecture of Phase I. (The prototype implementation focuses on the OHDSI paradigm, as the more technically challenging audience: the other paradigms are currently supported by many of the INSPIRE member sites.)

Examples of the members of that type of audience are listed in each case. These are intended to be exemplary rather than comprehensive. This analysis was conducted mainly for the purposes of setting the stage for the work and communicating about who intended users are. As mentioned above, Phase I focused only on the first category: microdata users. While this breakdown is mainly useful in understanding the documentation and metadata requirements of users, it will also inform the way in which topics such as access control and non-disclosure are addressed.

C. Meeting Metadata and Documentation Requirements

As described in the preceding section, users of the OHDSI applications will already have the information needed to perform their analyses – they do this today, and the INSPIRE Hub would operate within that paradigm. For those classes of users which will perform analyses, but who may not be using the OHDSI tools, other ways of accessing documentation and metadata will be needed.

One of the commonly used standards for this type of metadata and documentation among researchers within the INSPIRE community is the *DDI Codebook* standard, which has come into common use in Africa

through the work of the International Household Survey Network, and which was adopted by INDEPTH and subsequent projects. To show how non-OHDSI expressions of metadata and documentation can be provided by the INSPIRE Hub, we will use this standard as an example off-ramp.

In this format, all of the documentation and metadata is expressed as a single XML file which is used in combination with an ASCII expression of the data set – exactly the way in which data would be received from the INSPIRE Hub when a cohort (in OMOP CDM terms) is expressed as a SQL query and submitted (the examples in the *Book of OHDSI* use the R libraries to perform this function, but it could be done with any tools which support this ubiquitous relational query language. For the INSPIRE prototype, existing tools for working with POSTGRES SQL were employed.)

Note that the OMOP CDM provides two types of cohorts: rule-based and probabilistic. We will only address the first of these. Further investigation into probabilistic cohort definition will be needed if we wish to support this outside of existing support within the OHDSI tools framework.

The cohort expresses which fields from which tables will be included in the result set, and also dictates their ordering. Further, the ODHSI cohort definition defines exactly which concepts are to be included in the result set. The details of these concepts can be found in the OMOP CDM.

The *DDI Codebook* model gives us several levels at which metadata can be held, of which two are of primary concern: the “study level” information (documentation about provenance, methodology, and other information applying to a data set as a whole) and “variable level” information, which includes metadata at a granular level for operations on the ASCII data set by both the human users and programs (e.g., for transformations).

The OMOP CDM cohort definition expressed as an SQL query provides us the structure of the data and gives us sufficient information to programmatically describe the result set in terms of the variables contained and their arrangement into records. The vocabulary tables in the OMOP CDM give us sufficient information to describe the values of those variables, and links to the concepts from which they are taken.

The documentation at the study level is more problematic: in this scenario, the data is potentially coming from a wide range of sources, and these may employ different means of capturing the data. Where the assumption in OHDSI-based research is that the data are coming from a variety of clinical sources, the assumption behind *DDI Codebook* is that the data are the result of the administration of a survey or have been collected at a single point in time from an administrative register and rendered into a static file. The “Study” constrict in DDI describes this event. While OMOP CDM does provide a means of describing measurements which have been captured using questionnaires, there is no assumption that all of the data in a data set will have been captured from the *same* administration of a questionnaire.

In order to provide support for the documentation and metadata at this level to users operating outside of the OHDSI paradigm, the “Study” must be defined: for our purposes, it is the execution of the OMOP cohort itself.

Once the result set is in hand, it them becomes possible to know the exact list of providers for the data, and even which specific mechanisms were used to record them (which questionnaire, for example). This

information can then be used to assemble a description of the overall contents of the data at a summary level and included in the *DDI Codebook* study-level metadata.

In looking at the results of this approach, it became clear that the specific methods and practices at each of the data provider sites may well differ, and that it would be necessary to track them alongside the data they supplied. The OMOP CDM provides for this, through the use of their METADATA table, in combination with additional FACT_RELATIONSHIP table. Although use of these to track which data provider had contributed any given data point to the INSPIRE Hub would not be understood by generic users of the OHDSI tools, this would have no particular impact: the reason for capturing this information would be strictly to support non-OHDSI users, and so would be specific to the INSPIRE framework. The fact that it represents a "non-standard" use of the typical OHDSI conventions is thus immaterial for users within that framework.

If the data providers can be known, documentary information about their methods and practices can be included in the *DDI Codebook* instance as needed, as described in the section above. These would be stored external to the OMOP CDM instance, which could provide links as appropriate within the METADATA table of the CDM instance.

It should be noted that this mechanism was explored but not prototyped within Phase I of the ALPHA project. Such an approach does appear to be feasible, however, and would make it possible to support the range of users of tools which addressed by the DDI specifications, including STATA, SAS, SPSS, and some others. Such an approach would involve the production of the ASCII data result-set, and the DDI metadata file realized as part of a process which could be developed on top of the INSPIRE OMOP CDM instance.

VII. MANAGEMENT

A. Background

There are a number of areas which must be considered by the INSPIRE Hub. Although not itself a data repository or archive – it acts more as a service for harmonizing data across disparate sources – there are still aspects of data management which cannot be ignored. This section addresses those functions in terms of the support for data provision and use described above.

These topics include access control and management of users, disclosure risk control, support for data discovery and citation, quality assurance, and requirements around audit trail and data revision. None of these topics is explored in detail, but the general approaches considered are described. In every case, further work is needed to determine the best path forward in addressing this variety of concerns.

It should be noted that the "immutable cohort store" mentioned in the high-level overview is a key component in addressing many of the issues covered in this section. Part of the challenge faced by the INSPIRE Hub is to bridge between user communities, some of which are accustomed to working in a "file-based" paradigm toward which many of the existing practices (such as data cataloguing and citation and disclosure risk control) are oriented. Part of the challenge for the INSPIRE Hub is to produce a mechanism for bridging this gap, and the immutable cohort store is a key aspect of the architecture in this regard.

B. Access Control and User Management

Access to the data is intended to be provided to qualified researchers who have been determined to be trustworthy and legally allowed to access the data in the INSPIRE Hub. This process is not one which is primarily technical, but one which relies on established practices which can be dictated by existing practitioners who are members of the INSPIRE Network.

On a technical level, the policies determined by the INSPIRE Network must be enforceable, however. The mechanism for managing access at the level of the underlying database used by the INSPIRE OMOP CDM instance is proven and relies on existing relational technologies. However, the database access controls must be managed at a higher level, to determine levels of access which are driven by policy considerations rather than technical ones.

Data coming from some countries may not be legally available to researchers in another; licensing may determine whether access is permitted; and so on. While such requirements have not been explored in detail, it is significant that the OMOP CDM concept of cohorts is one which may be useful in this regard. Cohorts are essentially subsets of the overall data held in the instance, measured according to some criteria. Further, they can be combined: a cohort can be superimposed on another cohort, as is indeed the case in some types of clinical research.

This mechanism could be leveraged to reflect “cohorts” which reflect not the focus of research, but the limitations of access profiles. Each set of access conditions could be defined as a cohort, and the needed conditions associated with the access granted to any user. When the data was being requested, these cohorts could be used to refine the data available to that user through the Hub.

This approach has not been fully explored, but it offers a promising avenue to explore. Given that existing tools in the OHDSI ecosystem support the definition and use of cohorts in combination, it represents a powerful existing mechanism which has the appropriate nuance to reflect the limitations imposed by real-world considerations. Further, it is typically the data about individuals which must be controlled, and the cohort mechanism in the OMOP CDM is primarily designed for the subsetting of exactly this type of information. It is typically not the case that metadata and information of other kinds is subject to similar controls.

C. Anonymisation and Disclosure Risk Control

Privacy protection can be challenging, because it is necessary to address it at several different levels. The initial data submitted by the producer must be pseudo-anonymised, but the microdata and tabulations of it may also be disclosive as a result of how they are combined. These issues require that both the data producers and the Hub work together to control disclosure risk.

There is no way to guarantee in absolute terms that data will be non-disclosive. There are, however, accepted approaches, guidelines, and standards which can be used to define a best practice which provides an acceptable level of risk. We anticipate that not only will we require and validate compliance with such practices but will also need to support data producers in effectively implementing them. Fortunately, there are in many cases existing tools for preparing safe data.

One of the tools in our arsenal is the distinction which can be made between public use and scientific use files. We will in some cases need to check that a potential user is a bona fide researcher before allowing them access, because the potential disclosure risk is too high for general access to be allowed.

Examples of how data catalogues handle this distinction exist, and can be imitated (e.g., LSHTM, IHSN NADA Catalogue, etc.)

The approach to controlling disclosure risk can be understood as existing on three levels:

1. *De-identification (pseudo-anonymisation)* – name, phone numbers, national id numbers, etc.
   This is handled by data producer who is submitting data to the Hub - we can check/assist with well-established techniques. (Technical project lead Chifundo Kanjala and others have written a relevant paper on the topic: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3648430)

2. *Statistical Disclosure Control* – microdata
   This is an examination of the assembled variables, to make sure that their combinations are non-disclosive. There are standard guidelines, which require customization (esp. for longitudinal data). This is a function of the Hub for any data containing non-public variables. (See also Chifundo's paper referenced above.)

3. *Statistical Disclosure Control* – aggregates
   Tabulations and aggregate cubes must be checked to make sure that sample sizes and other potentially disclosive aspects are identified and corrected. There are methods (cell suppression, etc.) and tools for doing this, but this will be the responsibility of the Hub for any aggregations it distributes. We should make tools available to users who will be doing their own tabulations from our microdata. There are guidelines in this area from UN and other international organizations (e.g., the World Bank) which we should follow and make available. [Citation: https://www.ihsn.org/anonymization - covers both microdata and others (ESSNet: http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf), Matthias' book: https://www.springer.com/gp/book/9783319502700]

There is a further issue here: we need to give providers of data to the Hub confidence that their data will be non-disclosive and safe to share. Since we know that these organizations may not have these skills, we need to validate that data is in fact safe, and to support international best practice to empower the producers (tools, guidelines, etc.).

Chifundo Kanjala is currently working with Dr. Matthias Templ from Zurich University of Applied Sciences putting together a set of requirements for anonymising HDSS longitudinal data, and this will provide an addition basis for future work in this area.

It should be noted that part of the challenge is that dynamically queried "data sets" (that is, result sets described using the OMOP CDM cohort) can be difficult to anonymise, as this function is best performed as part of the process which returns the result set. While not unheard-of, such "run time" anonymization is less common than the same function performed on static files. To the greatest extent possible, data holdings should be profiled for access (see above) such that no combination of them would produce a potentially disclosive result set. It may also be possible to have pre-determined cohorts for use which have already been thoroughly vetted, and which are known to present no risk of disclosure (see the section on citation below). Further work is clearly required in this area.

D. Data Discovery and Citation – Pre-Defined "Data Sets" and Archival Requirements

The planned breadth of the INSPIRE Hub is broad, and there is a requirement that researchers be able to discover what is contained within it, and also – once data from the Hub has been used in research – to be able to cite it. Both of these requirements exist in a context where the data held within the Hub will be changing as new data is added, and as existing data is revised.

Discovery of data is a challenge which can be met with existing standards and approaches, so long as one basic requirement is met: almost all existing cataloguing and discovery mechanism have a presumption that data is contained in "sets" which – while they may expand along familiar lines – can have their contents described according to the fields which they comprise. As designed, the INSPIRE Hub will not meet this requirement: as new data is added, the coverage of that data will also expand. A mechanism for packaging data into some form of "data set" (as mentioned above) is required.

For the purposes of data citation, the same is broadly true: there are good standards for citing data (DataCite and similar approaches are common and use the popular DOI mechanism). There is no need for INSPIRE to invent its own scheme of data citation. However, not only must cited data be contained in a citable "data set", it must also contain only that data which existed at the time it was used for whatever purpose the citation results from.

Given that data can be updated and revised, a mechanism is needed for recreating exactly the data set used at a specific point in time, and the citation must be made to that exact result set, not to a similar one with revised content.

The OMOP CDM provides a useful basis for addressing these needs, and in the INSPIRE architecture these will be employed in the form of the immutable cohort store. The basic packaging mechanism in the OMOP CDM is the cohort: it determines exactly the set of data to be included in a result set at the time in which it is executed. This mechanism gives us the basis for addressing both the problems of data discovery and those of data citation.

The idea is that if a cohort can be preserved by the Hub, and the exact time at which it was executed can be associated with the cohort definition, then the result set it produced can be recreated. This relies on the fact the OMOP CDM provides support not only for the observation periods for which data is relevant (and by which inclusion and exclusion in a cohort is described), but also the time at which any given data point was put into the system. Using this set of information, it is possible to re-execute a query to produce the exact result set which was obtained at an historical point in time. (The combinatorial nature of OMOP CDM cohorts was described above for use in access control – a similar use of that mechanism here also suggests itself.) It is this type of time-bound cohort execution to which a data citation can be made, and a DOI attached. This is one function of the immutable cohort store.

The problems around data discovery are less demanding but can benefit from the use of a similar mechanism. It would be possible to leverage common schemes for supporting data discovery such as Schema.org and the W3C DCAT vocabulary, but these make the assumption that data is organized into to topically coherent "data sets". Again, a representative set of pre-defined cohorts could be designed and stored, although the time of execution needed for data citation would not be required.

We have already described how such cohorts could be documented (at both a detailed and a general level, as for non-OHDSI users) using standards such as DDI Codebook. This exact metadata and documentation is also that required to support common discovery standards such as those mentioned. (Mappings from the DDI standards to many common discovery formats already exist.) If the immutable cohort store can provide access to the contents of the INSPIRE Hub in a form which behaves like a file-based "data set," then it can also support the discovery of data within any of the typical data catalogues, or using any of the functionality for data search based on standards such as Schema.org.

Further, the described functionality could be used to support many aspects of audit trail and archival requirements, although not extending to cover the needs of data preservation. Given the intended use of the INSPIRE Hub, this limitation is not seen as meaningful.

E. Quality Control

It is assumed that the INSPIRE Hub will work with data providers to ensure a satisfactory level of quality in the harmonized data to which it provides access. Such up-front quality control is not addressed within the Hub architecture but is assumed to be taking place within the context of data acquisition. (Existing INSPIRE partners already provide data of high quality.)

However, the ETL processes and the mapping of coding to the standard OMOP CDM vocabularies can raise the need for further quality controls and ensuring that the minimum technical requirements of the system are met is also important. For these purposes, it is anticipated that existing OHDSI reporting tools (notably the Data Quality Dashboard) would provide a good foundation, supplemented by whatever other management processes would need to be put in place by INSPIRE system administrators to guarantee integrity and correct functioning of the Hub.

VIII. LOOKING FORWARD

A. Overview

Many areas discussed in this document thus far indicate that there is more work to be done and given that INSPIRE Phase I was limited in scope, this comes as no surprise. The project established that the OMOP CDM and the accompanying OHDSI tools provide a sufficient framework for providing access to harmonized clinical and population data. Many areas remain to be explored, however. The following section mentions some of the areas which are proposed for further work.

B. Distributed Access

The OHDSI community has a concept of "networked research" which is intriguing, involving data structured according to the OMOP CDM coming from a number of different sources. Such data is often limited to aggregate data, however, which may not be enough to meet the requirements of all users. This approach, and some other ideas in this area, remain to be explored.

The issues to working with data from across several different sources are many, and there are existing efforts to address them. In some cases, issues stem from a difference in data structures, coding, and semantics. In other cases, data is not legally allowed to be processed outside the physical confines of the country in which it was collected. The potential for unacceptably high disclosure risk is high.

Members of the INSPIRE Network have considered this problem, and some approaches have been discussed. These will require additional work to fully explore.

The known approaches include:

1. Hold links to other data within the INSPIRE framework but require users to access and harmonize that data themselves (this is the approach used by many archives and data catalogues, such as the IHSN Nada catalogue). Relevant sources of data would be associated with the various data held within the Hub, but only the links referencing the source would result from queries against the INSPIRE Hub.

2. Do not hold data but allow run-time transformation by resolving the links held in the Hub and applying pre-designed mappings into the OMOP CDM at run-time.

3. Adopt the OHDSI "Network Research" approach with other OMOP CDM instances (covers aggregates only). This would require that any participating data providers have a compatible implementation of the OMOP CDM.

C. Increased Data Coverage

So far, the INSPIRE Hub has looked at the combination of population and clinical data. Other data – especially in the age of COVID-19 – could quite usefully be added to the resource for easy combination with expected current holdings. Such expansion could include genomic data, for example.

For any given set of data sources, on-ramps and off-ramps will need to be identified and constructed to support the inclusion of the data. The OMOP CDM provides a very flexible basis for this work, but each new type of data will require identification of sources, analysis of what ETLs and mappings are needed, and the design and production of these. If new user audiences are involved, work to map the INSPIRE Hub holdings into useful forms to support additional off-ramps will also be needed.

D. Collaboration with International and Regional Initiatives

In looking toward the future, it is clear that developments within the ODHSI community may offer new tools and approaches to be leveraged within INSPIRE. Further, there are many activities at the regional and global level which are in line with what INSPIRE is aiming to do. Among these are the different open science clouds which are emerging. Of most importance to INSPIRE, perhaps, is the African Open Science Platform. In other regions of the world, similar developments are occurring – the European Open Science Cloud is one, and there are others in China and Canada. There is even a community emerging around the idea of a Global Open Science Cloud.

At a different level, organizations such as UbuntuNET are organized as regional NRENs, and these offer an environment in which the INSPIRE Hub could function as a data-focused application within a related set of research-oriented ones.

It is anticipated moving forward that the prototyping and design work undertaken in Phase I of the INSPIRE Hub could adopt or contribute to many of these different efforts, and further relationships with these groups could be usefully pursued.