Health Data Exchange (HDX): A GLOBAL Open Data, Software, and Learning Community Resource for R & D Acceleration and Dissemination

> Javed Mostafa Dean & Professor University of Toronto ischool.utoronto.ca

OUTLINE

Introduction Why data exchange? Major goals System architecture Sustainability FAIR Metadata Annotation Summary





INTRODUCTION

In recent times the volume of digital data generated has skyrocketed to about 2.2 Exabytes per year

1 Exabyte = 1 Billion Gigabyte, 1 GB is about 4500 books, 200 pages each

3

Of the total volume about 30% of it is healthcare data – about 0.7 Exabyte ~ 1 Billion Gigabyte by 2025 (at current growth rate)

Health data generation grows at a faster pace than any other major sectors (e.g., manufacturing, financial, media and entertainment)

But, due to many security, IP, and bureaucratic reasons accessing and mining healthcare data is exceedingly difficult (even inside an organization)

 $https://www.rbccm.com/en/gib/healthcare/episode/the_healthcare_data_explosion$



WHY HEALTH DATA EXCHANGE (HDX)? REASON 1

4

Data analytics, machine learning, and a majority of research in biomedical and health sciences

Require ACCESS TO DATA and produce DATA





WHY HEALTH DATA EXCHANGE (HDX)? REASON 2

5

Very hard to gain access to data in a

- Consistent,
- Persistent,
- Standardized,
- Regulatory-governed, and
- Sustainable way





WHY HEALTH DATA EXCHANGE (HDX)? REASON 3

6

Post-analysis or post-research, data still has value

- Publication/dissemination
- Replication
- Training/pedagogy
- Follow-up or scale-up work
- Community- or team-building



MAJOR GOALS OF HDX

A phased design, development, and sustainable approach



AREAS OF FOCUS

SUPPORT RESEARCHERS

SUSTAINABLE PLATFORM

- Findability of data
- Accessing data
- Using data
 - Support software pipeline and scalable computing
 - Support secure and audittrackable use
 - Support dissemination of results
- Contribution of data from researchers
- Updating and retirement of data

- Community-building
 - Linking to publications
 - Linking publications
 - Linking researchers
- Training and pedagogy
- Partnerships with publishers, non- and for-profit organizations

EXAMPLE OF PEER RESOURCES: NCBI

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

Download

Transfer NCBI data to your

computer

About the NCBI | Mission | Organization | NCBI News & Blog



Welcome to NCBI

Deposit data or manuscripts into NCBI databases



Develop

Use NCBI APIs and code libraries to build applications Identify an NCBI tool for your data analysis task

Analyze





UNIVERSITY OF TORONTO FACULTY OF INFORMATION



https://www.ncbi.nlm.nih.gov/



Find help documents, attend a class or watch a tutorial



Research

Explore NCBI research and

collaborative projects

9

EXAMPLE OF PEER RESOURCES: NATIONAL COVID COHORT COLLABORATIVE OR N3C

ties /	Et al. an - Frankeling Right for the Annotability, associated insertion, map institut, Bill Annie, and Anniel						1
000 000 00 000 0 0 0000				and the second			÷
-	10 Tes	-		Sec. Star	Distance of the local	-	
Investigation Transients	 Mage of patient country of approximations (see) 	to par in	A. 10.10	Rear - Model with the columns and space. And with the columns of a rear of the last with the fact having of a start of a space with the fact having of a start of a space of a start space.	March 1990		
Managama	D to how how how	All and Associate	N	No implemente la colte cita militaria e citale electrica en las primero della constante e de las primero della Compacta en las actuales ante Compacta en las alte	at many cost of		,
Restator Texplete 11 Decision	Contact Sector correspondences	1.04	Pa. or 1, 107	The section is an and is and that MMP when is advance automation topics for the sector automation similar in same interpretation when it same interpretation is an	Ref. (and the local distance of the local di		
	Contraction Pagente for address states and generation	A	Beller 1.00	The day formula and definition of the se- state and an ending with the following forms: manifest with the manufacture of the S	Mr. Protectures Research Proper		
historiate. •	O Granet a course	8111 Tarra	No. 14 (0, 10)	The second secon	an construction in the second		
Tale Tergens II	C Statistical Alignments	A	Res. 14 (0. 107)	63 metric title annum Annumenter, för storer galannet och metrick för äverationen annungen annungen för annungen för annungen för annungen storer	March Street		
	C Price Sector State	A	Tel. 10.107		March Contractor		

	-				-	-	
	-	-					
	RFB-1 Decyl Administration	contract in	Jane Holgen	1000	Anatat	Date of the local division of the local divi	10.000
	Carl Contract (Carl Contract)	Manuf Lane	10.000	10.00	Anatat	And in case	-
Sale (sectorized as 100 million	41.41.585 (st.) + 70.4418	Table Tool I	Ann Nown	Contraction in the local division of the loc	Available	And in case of	10.000
att termine 1	Carlos and a second	Married Voters		Aug. 11, 1010	Annal	Anatal	10.000
	A fair and a special sector of Mag 1078	1410	And in the	10.100	Annal	And the lot of the lot	10.000
Antiplate In	Real and single language	88.948	Ann Terror	10.00	house	And in case of the local division of the loc	10.000
Condition in Will Review 17	Contact - 18	Datase (M	Beyon Inc.	10.00	house	August	No. of Concession, Name
Analysis Investor, 17	Contact - M	Dates #	Beyons Inter	Arr. 2. 1998	house	August	-
Autophthese 11	Charlos Maurato	Contract.	Brown Inc.	\$1.0.00	Nume	August	-
State State	Conversion in the second state of the	forwards		22.2	Realist	And in case	-
	C time marks	to down reply	Acres 10,000 (10)	101.000	Trans.	functions for	-
	C tanca	Transa.	Revelle	Acr. (8, 102)	(house)	And in case	No. of Concession, Name
	to may and mean	the Year	Revenues	147.00	Reality	And in case	-
	Tantan diana array again	Francis of Non-	Bergarin Array	Aug. 11, 2020	Report	Assisting Will	-
test of	Contra Co	and the second	three-balance	Sec. 10, 1000	Report	Analysis	-
An er sonser sonsen	E RO MAR	Responses to the local division of the local	Tel Bernet	\$1.0.00 	(install	And in case of the local division of the loc	-
ANT. 1	Another environments and an environment	and manys.	Area Telepon	10.00	Realized	function for	

N3C Knowledge Store

Access and secure sharing platform for analysis Browse curated medical code sets, including peertemplates and derived or ingested publicly available review validation information. datasets.

Bostines	# hors & Bestellus			
bes-Leving Interters File				August 100
-			100	
E faitheast and	Agent allows and indefinition th			
E transition	s			
E fanty hate	w Toerials		-	
E AL NUMBER	Res 1. Krage is divide all spec	*		
E DEP Los Sec	-	*		
Contractory	an an annan ann ann Agara barann an ann	*		
Auto Sala Sala	na Talan mangal ini ing Talan di Basari i selatara	*		
Contraction of the	fee	*		
Entra More	Comp to a SACP terms the last frequency and the Taylor (2010) TO 2019 When a HTAC spectra automation of the SaCP terms of the International Conference on	*	•	
E tot familie	No Metrical In Sectorial Construction and Sector And Construction (MICP Value)	*		
E nations		*		

Data Catalog

Browse the most commonly accessed data tables as well as notional data for learning.

Filler Roles	(i) with Contracting Matter	Retro Springer -					
	-	Martine state (second provided periods), we have in tradinate a latent of the state (second second secon					
	What have subject and a phononal and then do indicate a between of the subject from the	I I have bare officerstrikers I have been I be b					
	Marchine Berne or patie. Apriles, Spollar and countain fails	B terret to:					
at.	Figure International with the INCP International Index Section (International)	What is a 'value?' and a 'shanstyna' and how do I					
	Sample cole aprillage for my carry a	obtain a dataset of my cohort from the DMOP					
ad anation.	form a 1969 failer faile failer	(append)					
	Non-th-Tabelity Properties was to be \$207	Based on the DRDB definition. A calculation of a proving who calculates on more inclusion othering to phenotypel Ter a duration of time.					
	 Now to december a concept and 	A detail containing periods is poor othert of mission on the detail from your own					
) with line log	time is perform a circuit mone of a concept and	contermation method. Since you have defined your cohort and have the color to the out- we highly encourage you to instea a mate-mate temptore of the color? Methods code to that others con color your color/for additional research.					
	What are the considerations related to been heterogeneity tensor unter?						
	Intending repeated ecceptions will rank attach and not like	We can have determine the phase available that dense objects the factories and interesting and 2014 beth based on a colory. 2016 Automatics, fit can there implete phase early will apply target bases in 1 fieldness part offset in 2018 phase about one of the phase is based on the phase of the colored in the phase. If early the colored additional is balance in 2018 and the colored in the phase of colored additional in 2018, and the colored colored in the phase of the color additional to a solar.					
	West date "publicking" a concept of reser-						
	The or deally when have						
	Considerations when using its shaft ratio						
	The other are says in the ancient refracted?	9 Nor Data Gitzenson III (2 - Y)					
	himse in particular a consideration internet of a converged and	Carlos has Review					
	Dealing can fandar a NE						
	Genhauss Meanth across angless?	C					
		L. N.L.L.					
	mmiini						
	munum						

es as The N3C Community Notes application provides a place for users to asynchronously share documentation, tips, and links to other resources with the N3C community.

NCATS National COVID Cohort Collaborative

Importantly, since the shared data could be partly identifiable, the data is stored in a secure platform, with strictly regulated usage. Researchers who meet the requirements (<u>https://ncats.nih.gov/n3c/about/applying-for-access</u>) can apply for access to the data, for use only for COVID research and response planning.



DETAIL VIEW OF HEALTH DATA SHARING: USE CASE N3C





GOALS FOR HDX

FINDABILITY

 Resources need to be found quickly and easily

ACCESSIBILITY

•

- Resources need to be searchable based on standardized terminologies and common words
- Retrieved results should be easily browsable and understandable

- USE
- Results should be downloadable
- Tools should allow the user to analyze the data using scalable compute resource
- Output should be easily storable, recoverable, and sharable
- Allow contribution of new data
- Retirement of access and data

COMMUNITY

Support finding the contributors /producers of the data and communicating with them

•

•

- Support tools for identifying teams and resources
 - Provide tools for conducting meetings, access relevant resources, and store outcomes of meetings/
- Platform for collaborations

TRAINING/ LEARNING

- Support tools for accessing and using data for teaching purposes
- Provide documentation and training materials and link appropriate data to the learning resources
- Provide a way to link external learning resources to internal data



٠

UNIVERSITY OF TORONTO FACULTY OF INFORMATION



Users can access or use data, or they can engage with a community Users can use software on a data set using builtin libraries of analytics and visualization tools Data set layer allows simply browsing existing data, uploading new data, or modifying or removing user-owned data Application Programming Interface layer which allows access to data using common protocols (e.g., HL7 FHIR)

HDX ARCHITECTURE: DATA CATALOG I



HDX ARCHITECTURE: DATA CATALOG II



HDX ARCHITECTURE: SOFTWARE PIPELINE I



HDX ARCHITECTURE: COMMUNITY







SUSTAINABILITY I

USER COMMUNITY

- Provide organizational and domain-based directories for the user community to find collaborators
- Support discovery of other researchers and experts who have association with data sets (produced the data or used the same data set before)
- Allow users to identify relationships between domains and members of the research/expert community
- Support functions such as posting of blogs/comments and securely message members of the community



SUSTAINABILITY II

USER COMMUNITY

- Create "Gold Standards" for different standardized machine learning predictive models and classifiers
- Support "Friendly but Rigorous Competitions" on an annual basis to promote awareness and develop closer bonds among the users of data and software
 - A nice model is US NIST TReC Data Sets and the user community which NIST nurtures

Text REtrieval Conference (TREC)

...to encourage research in information retrieval from large text collections.



trec.nist.gov



SUSTAINABILITY III

PARTNERSHIPS & BUSINESS MODEL

- Partner with research centers/universities, publishers, NGOs, non-profit and for-profit research organizations, and networks of current data hubs (e.g., NCBI)
- Shared IPs with partners on advanced data access and manipulation tools
- Establish a non-profit status for HDX but have a solid business model



FINDABILITY, ACCESSIBILITY, INTEROPERABILITY, AND REUSABILITY (FAIR)

Developed by group of scientists and practitioners from organizations worldwide [2016]*

Easy to appreciate if you consider the original emphasis was on "machine" discovery, access, and use of data

- Finding and discovery: Improving visibility of content using identifiers
- Accessibility: Terminology and ontology standardization
- Interoperability: Semantic integration
- Reusability: Securing data and use of secure protocols





FAIRIFICATION PROCESS



23

The three components of the FAIRification Framework: the conceptual FAIRification Process, the FAIRification Template covering all aspects of FAIRification and the FAIRification Workplan as a single tailored implementation guide.

https://www.nature.com/articles/s41597-023-02167-2

SUMMARY

- Data, software, and people information need to be aggregated on a single platform
- Findability, accessibility, manipulability, and sociability are key dimensions
- Must support pipelining of software to promote usage and replication
- Sustainability can be ensured through community and partnership building
- Must support FAIR principles



THANK YOU

Javed Mostafa dr.javedm@utoronto.ca ischool.utoronto.ca