Data Sharing Spoke Initiative

2018 Annual Summit

of the Northeast Big Data Innovation Hub

Jane Greenberg, Director, Metadata Research Center
College of Computing and Informatics
Drexel University

IIS/BD Spokes/Award #1636788

Overview

1. Context

2. Spoke Initiative-update

   "A Licensing Model and Ecosystem for Data Sharing"

3. Next steps

Team members

- Sam Madden, Lead PI, Massachusetts Institute of Technology
- Carsten Binnig, PI, Brown University
- Sam Grabus, grad. RA, Drexel University
- Jane Greenberg, PI, Drexel University
- Hongwei Lu, grad. RA, Drexel University
- Famien Koko, grad. RA, MIT
- Tim Kraska, PI, Brown University
- Danny Weitzner, PI, MIT

PROJECT PAGE: http://cci.drexel.edu/mrc/research/a-licensing-model-and-ecosystem-for-data-sharing

IIS/BD Spokes/Award #1636788

How open data on agriculture & nutrition can solve world hunger

Give Up Your Data to Cure Disease

By DAVID B. AGUS   FEB. 6, 2016

THE CURE FOR CANCER IS DATA— MOUNTAINS OF DATA

WIRED
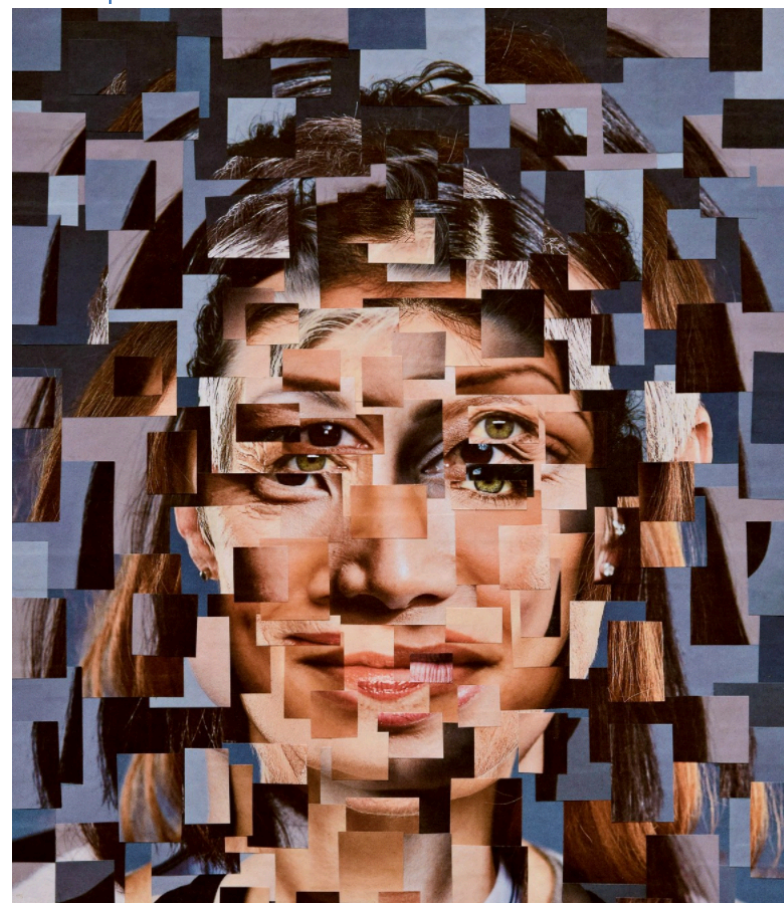
Has anyone deposited data in a repository?

- *Open*

- *Restricted*

- *Don't know…*

  - *Haven't but thought about it…*

Has anyone here shared research data with a colleague?

I did!!

It helped me get tenure…

Has anyone here ever thought…

- *WOW, if only we could get DATA…*[HEALTH RECORDS] [PURCHASE BEHAVIOR] [ENERGY CONSUMPTION] …we could test our algorithm, and conduct seriously robust research that has a real impact

- *BUT… I cant because of…*
  - *Legal issues…*
  - *Privacy…*
  - *Policies*

Data sharing barriers

Policy	Licensing, agreements	
- Complex regulations governing use of data in different domains - Data lifecycle – data…living thing ~ *Do not want to loose control over data downstream* ~ *What if data is redacted?*	"Creative commons" (CC, CC0, etc.) does not address need	**Rights, privacy**
		Concerns over sensitive information (e.g., PII)
	Security	**Incentives**
	Technical and systematic aspects	Why would someone go to all the effort to share their valuable data?

Still, merit in sharing

No sharing without a legal agreement

Involves lawyers to create individual agreement!

Spokes and rings

Co-Chairs
Jane Greenberg, Drexel
Sam Madden, MIT
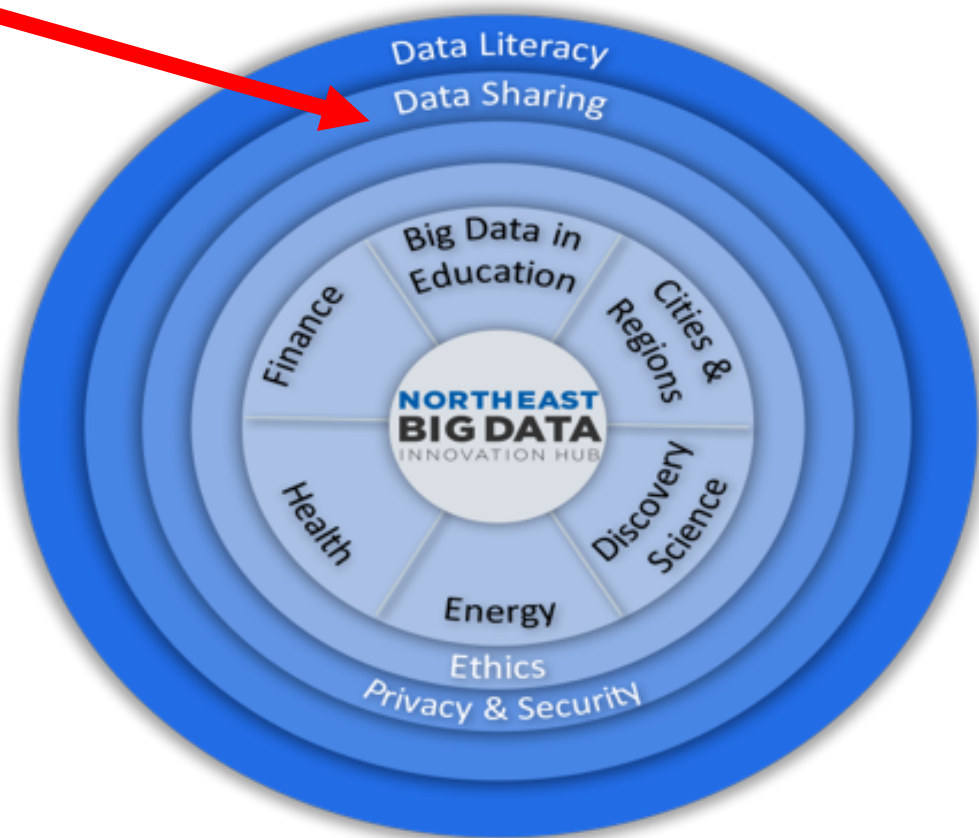
Spoke Project: A Licensing Model and Ecosystem for Data Sharing

1. Licensing Framework / Generator

2. Data-Sharing Platform (Enforce Licenses)

    - DataHub

    - ShareDB

3. Metadata (Search Licenses and Data)

- Principle: Solve the 80% case!

Enabling Seamless Data Sharing in Industry and Academia (Fall 2017)

Heard from the trenches…

- Collect agreements
- Build a trusted platform
- Good metadata!

Progress

1. Metadata infrastructure and life-cycle management

2. Prototype system – DBShare

3. Algorithmic related work

Licenses Analysis
(Sam Grabus, CCI/Drexel)

High-level Categories

General:
attributes relating to the project and the agreement itself

e.g., Description of the data, Definition of terms

Privacy & Protection:
the protection of sensitive information and security

e.g., Individual identifiers removed prior to transfer, Encryption

Access:
who and how contact may be made with the data

e.g., Who has access, Method of access (approved hardware or software)

Responsibility:
legal, financial, ownership, and rights management pertaining to the data

e.g., Indemnity clause, Establishment of data ownership

Compliance:
ensuring fulfilment of agreement terms

e.g., Third party compliance with contract, Background checks for personnel

Data Handling:
specifics of permissible interactions with the data

e.g., Publication of data, Conditions for Termination

Privacy & Protection

Sensitive Information

Regulations	Preparing data	Access
• Regulation used to define sensitive data (e.g., HIPAA, FERPA, etc.) • Compliance with federal/state/international data protection laws and regulations	• Identification of confidential/special categories of information (e.g., pii, proprietary) • Individual identifiers removed/anonymized prior to transfer	• Who has access to pii/confidential data • Who has access to proprietary information
Privacy	**Avoiding re-identification**	**Exceptions**
• Anonymization of data • Confidentiality and safeguarding of PII/sensitive data • Removal/nondisclosure of company/personnel identification in materials and publications • No contact with data subjects	• No direct/indirect re-identification • Statistical cell size (how many people, in aggregated form, can be released in groups) • Merging data with other sets (e.g., allowed with aggregated data—not in any way that will re-identify)	• Exceptions to confidentiality • Conditions of proprietary information disclosure • Conditions of pii disclosure (who, what, and for what purpose?) • Limitations on obligations if data becomes public • Limitations on obligations if data is already known prior to agreement • Limitations on obligations if data given by 3rd party without restriction

Security

• Sharing non-confidential data • Password protection/authentication of files • Encryption	• Security training for involved personnel • Establishing infrastructure to safeguard confidential data

Lay of the land: Agent, access/rights, + workflow

REQUIREMENTS	EXAMPLE METADATA STANDARDS
DATA PUBLICATION, DOMAIN DISCOVERY	
Persistent Identifiers	Product (Schema.org), DOI (Digital Object Identifiers), Handle system, OAIS (Open Archival Information System)
Domain specific schemes	Schema.org, RDA metadata directory or other resources
IDENTIFICATION/DESCRIPTION	
Personal Identifiable Information	Person (Schema.org) vCard (Virtual Business Card), VIAF (Virtual International Authority File), ORCID (Open Researcher and Contributor ID)
Organization profile	Organization (Schema.org), ORCID, NAF (Name Authority File), EAC (Encoded Archival Context) for Organizational Bodies
Attribution	Same as PII
LICENSING AND USE	
Access	MODS, The Recommended Practice Access and License Indicators (NISO RP-22-2015)
Restriction on Use	Embargos and Leases (Project HYDRA), PCDM (Portland Common Data Model: Rights Extension), METS, PREMIS (Preservation Metadata Data Dictionary)
Training/user requirements	Technical metadata, operational (see 'Technical Format' and 'Restriction on Use')
Technical format	Accessibility (Schema.org), W3C MS Global Access for All (AfA) Information Model Data Element Specification, PREMIS
Privacy	EHR (Electronic Health Records)
LIFE-CYCLE MANAGEMENT	
Workflow	Protocols found via scientific research, such as Taverna and Kepler will aid this work.
Provenance	PROV-Model (Provenance Model, W3C), PREMIS
Accountability/Authenticity	PREMIS
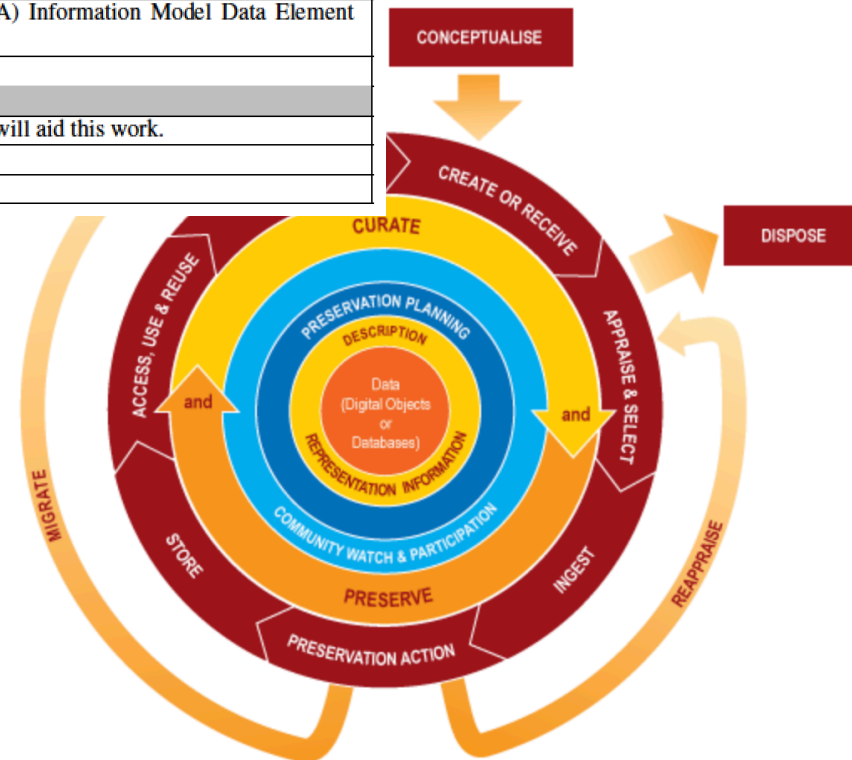
Rights related standards	Initiatives
■ Rights statements.org: http://rightsstatements.org/en/documentation/	■ Research Data Alliance Legal interoperability Interest GroupRDA/NISO Privacy Task Group RDA Metadata IG, WG Datasets licensing project: https://datasetlicencing.wordpress.com/
■ Mets: http://www.loc.gov/standards/rights/METSRights.xsd (rights declaration extension schema)	
■ Open Digital Rights Language (ODRL): https://www.w3.org/TR/odrl/, https://www.w3.org/ns/odrl/2	■ Dataverse data tags project
	■ Linked Content Coalition—LCC Rights Reference Model as part of the LCC Framework: http://www.linkedcontentcoalition.org
■ ONIX-PL for licensing terms: http://www.editeur.org/21/ONIX-PL/	■ FAIR-Findable, Accessible, Identifiable, and Reusable

Prototype system – ShareDB

Goal: Licensing Framework

Standard terms for researchers/data providers, lawyers, and compliance teams

- ☑ Controlled access
- ☐ Tracking of access
- ☑ Usage rights (e.g., publication, copying)
- ☐ Duration of use
- ☑ Warrantees of correctness/completeness/availability
- ☐ Other requirements

ShareDB

Guide to using ShareDB: Privacy Profiles

To create a new Privacy Profile and specify controls over your data set select 'Create New Privacy Profile'

To browse existing Privacy Profiles (made by you or other users) and add one to this data set select 'Add Existing Privacy Profile' and clic
desired Privacy Profile

Add Privacy Profiles

Create or change data privacy specifications for your data sets.

Create New Privacy Profile

Add Existing Privacy Profile

About   Documentation   GitHub Repo   API

ShareDB

Guide to using ShareDB: Privacy Profiles

Select desired privacy and security settings for your dataset. Once the Pro

Create new Privacy Profile for: testdata

Privacy Profile Name: | HIPAA PII Removed |

Regulations

- ☑ HIPAA ❓
- ☐ FERPA ❓

Privacy ❓

- ☐ PII Anonymized or Removed
- ☐ PII Anonymized
- ☑ PII Removed

Reidentification

- ☐ Use K-Anonymity ❓

K-size | Bucket Size for K |

Health Insurance
Portability and
Accountability Act

ShareDB

the table with th...

Once the Profile...

Apply Priva...

Profile name: h...
params: None

⊞ Base Tab...

testdata

Data Preview

Click edit for each data column to remove PII according to hipaa standards

IDENTIFICATION	FIRST_NAME	LAST_NAME	ADDRESS	PHONE_NUM	GENDER	SPECIES	RANDOM_SURVEY_ANSWER
edit	edit	edit	edit	edit	edit	edit	edit
1	Sam	Grabus	123 Sesame Street, Philadelphia, PA	867-5309	Female	Human	Yes
2	Jane	Greenberg	3141 Chestnut St, Philadelphia, PA 19104	555-5555	Female	Human	No
3	Kingman	Grabus	123 Sesame Street, Philadelphia, PA	867-5309	Male	Dog	Yes
4	Ted	Wark	103 Fayette St, Conshohocken, PA	123-5555	Male	Human	Yes
5	Morgi	Wark	103 Fayette St, Conshohocken, PA	123-5555	Male	Dog	No

Data Preview

Click edit for each da...

IDENTIFICATION								ANDOM_SU
edit								edit
1								s
2				Philadelphia, PA 19104				
3	Kingman	Grabus	123 Sesame Street, Philadelphia, PA	867-5309	Male	Dog	Yes	
4	Ted	Wark	103 Fayette St, Conshohocken, PA	123-5555	Male	Human	Yes	
5	Morgi	Wark	103 Fayette St, Conshohocken, PA	123-5555	Male	Dog	No	

Remove Column

✕

Click Delete to delete this column from the table

column name:
FIRST_NAME

Remove column

the table with the selected transformations applied

Once the Profile as been applied, you can preview created Privacy Profile View under 'Preview Dataset privacy settings'

Apply Privacy Profile To Tables

Profile name: hipaa pii removed

params: None

⊞ Base Tables   ➕

testdata

License applied ✔   **Apply Profile**

Preview 'testdata_privacy_profile_6' privacy settings

Tables & Views   Files   Cards

testdata_privacy_profile_6   ←

No description yet ✏️

Run Sentiment Analysis ▾

gender	random_survey_answer	identification	species
Female	Yes	1	Human
Female	No	2	Human
Male	Yes	3	Dog
Male	Yes	4	Human
Male	No	5	Dog
gender	**random_survey_answer**	**identification**	**species**

Algorithmic related work

Is this possible: Technology ⋈ Sharing Agreements

Technical

Access control & rights management

Expiration

Logging & auditing

Provenance/Finger printing

De-identification

"Noising"

Aggregation

Agreement Clauses

Controlled access (who & where)

Tracking of access

Usage rights (e.g., publication, copying)

Duration of use

Warrantees of correctness/completeness/

availability

Other requirements

Is this possible: Technology ⋈ Sharing Agreements

Technical

Access control & rights management

Expiration

Logging & auditing

Provenance/Finger printing

De-identification

"Noising"

Aggregation

Agreement Clauses

Controlled access (who & where)

Tracking of access

Usage rights (e.g., **publication, copying**)

Duration of use

Warrantees of correctness/completeness/availability

Other requirements

NLTK – parsing terms

- Set maximum keywords length: 5
  List top 1/5 of all the keywords

  ## Result:

  Keyword: research studies involving human subjects ,
  score: 20.4583333333
  Keyword: district assigned student identification numbers ,
  score: 18.8387650086
  Keyword: includes personally identifiable student information ,
  score: 17.6168132942
  Keyword: district initiated data research projects , score: 14.8577044025
  Keyword: support effective instructional practices , score: 13.0
  Keyword: personally identifiable information shared ,
  score: 11.3440860215
  Keyword: disclose personally identifiable information ,
  score: 11.1440860215
  Keyword: policy initiatives focused , score: 9.0
  Keyword: informing education policies , score: 9.0

-5	-4	-3	-2	-1	0	1	2	3	4	5
			educational	right	privacy	act	health	insurance	portability	accountability
applicable	federal	law	regulation	protecting	privacy	citizen	including	family		
	license	agreement	authorized	protect	privacy	individual	subject	nd	study	
				applicable	privacy	law				
consistent	federal	family	educational	right	privacy	act	department	designates	education	alliance
subject	federal	family	educational	right	privacy	act	authorized			
education	record	covered	family	educational	privacy	act	amended			
recipient	agent	subcontractor	violation	agreement	privacy	rule	security	rule	implementing	regulation
comply	applicable	state	local	security	privacy	law	extent	protective	individual	privacy
		data	security	protection	privacy					
information	identified	family	educational	right	privacy	act				
		de	identified	applicable	privacy	law				
				applicable	privacy	law	permit	data	provider	provide
				federal	privacy	act	requirement	apply	agreement	entered
shared	state	subjected	applicable	requirement	privacy	confidentiality				
resolved	permit	covered	entity	comply	privacy	rule				
time	covered	entity	comply	requirement	privacy	rule	hipaa			
		reference	agreement	section	privacy	rule	mean	section	amended	renumbered
					privacy	rule	extent	information	created	received
					privacy	rule	standard	privacy	individually	identifiable
					privacy	rule	include	person	qualifies	personal
tern	defined	agreement	meaning	term	privacy	rule				
set	accordance	term	agreement	hipaa	privacy	security	rule			
hipaa	regulation	promulgated	thereunder	governing	privacy	security	health	information		

Sentence with highest scores:

| privacy	protection	set				
applicable	privacy	law				
privacy	rule	standard	privacy	individually	identifiable
definition	set	privacy	rule			
data	security	protection	privacy			

Frequency from the most to the least:

Differentially Private Querying – Improving queries over data that don't expose private information.

Our approach automatically chooses an algorithm that will achieve a desired error while maximizing privacy.

For each algorithm in, A, create a model which can predict the epsilon to produce desired error rate, given dataset and workload -> select algorithm with highest predicted epsilon
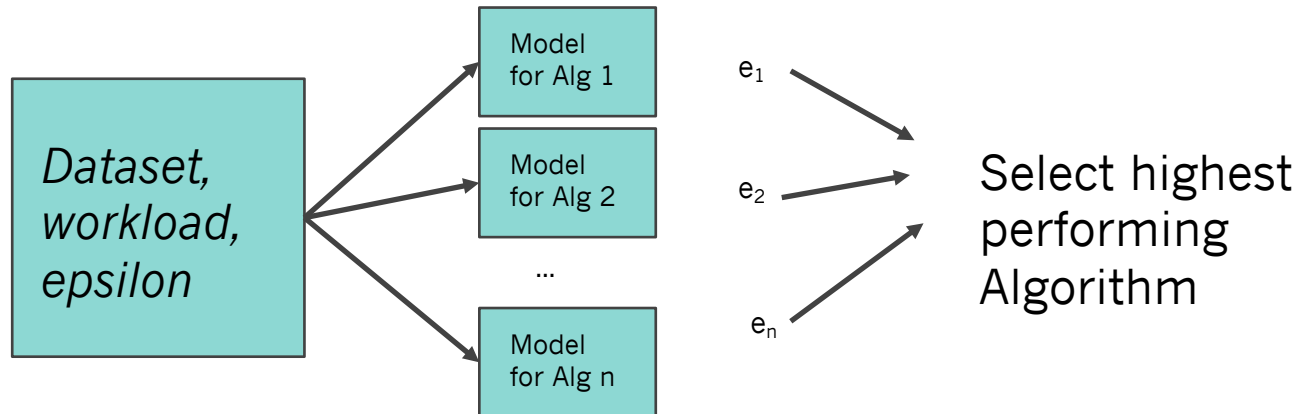
Conclusions and next steps

- Work underway, a lot of heavy lifting…
    - Mining licenses shows great diversity, but similarities
    - Metadata expertise
- Infrastructure to build on assisted with prototyping
- Continue to collect licenses
- Community building and connecting, IEEE-BDGMM, RDA – Research Data Alliance
- Webinar supported by NEBDIH (April?!!)
- Workshop-2, Drexel, September 2018

http://cci.drexel.edu/mrc/research/a-licensing-model-and-ecosystem-for-data-sharing

DREXEL UNIVERSITY
Metadata
Research Center
College of Computing & Informatics

ABOUT   RESEARCH   PUBLICATIONS   PEOPLE   NEWS & EVENTS   SF

CCI / Home / Research /

A Licensing Model and Ecosystem for Data Sharing

Project Summary

"A Licensing Model and Ecosystem for Data Sharing" is a spokes project led by researchers at Massachusetts Institute of Technology (MIT), Brown Uni as part of the Northeast Big Data Innovation Hub.

We are addressing data sharing challenges that are too frequently held up due legal matters, policies, privacy concerns, and other challenges that interf agreement.

Sharing of data sets can provide tremendous mutual benefits for industry, researchers, and nonprofit organizations. A major obstacle is that data often restrictions on how it can be used. Beyond open data protocols, many attempts to share relevant data sets between different stakeholders in industry a a large investment to make data sharing possible.

We are addressing these challenges by: 1) Creating a licensing model for data that facilitates sharing data that is not necessarily open or free between c Developing a prototype data sharing software platform, ShareDB that will enforce agreement terms and restrictions for the licenses developed, and (3) I relevant metadata that will accompany the datasets shared under the different licenses, making them easily searchable and interpretable.

"A Licensing Model and Ecosystem for Data Sharing" is also linked with the Northeast Data Sharing Group, comprising of many different stakeholders t widely accepted and usable in many application domains (e.g., health and finance).

Alternative ... repository deposition

By agreeing and submitting this license, you (the author(s) or copyright owner) grant to Drexel University Libraries the non-exclusive right to reproduce, translate (as defined below), and/or distribute your submission (including the abstract) in print and electronic format and in any medium.
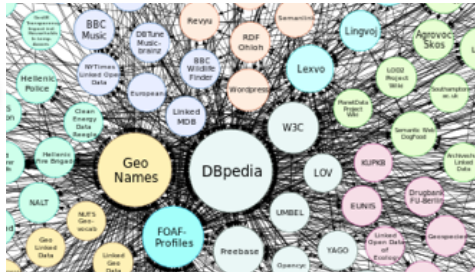
Open data

Closed data

Intel-Collaborative Cancer Cloud (CCC) (Dana-Farber, OHSU, Ontario Institute for Cancer Research (OICR))- *data*

Collaborative Genomics Cloud

(CGC) colocalizing massive genomics datasets) – *genomics sharing, identifying cancer causing mutation*

FICO score (Fair Isaac Corporation) – *credit score, risk*