GDSR FAIR

FAIR Data Standards in the Global Data Standards Repository

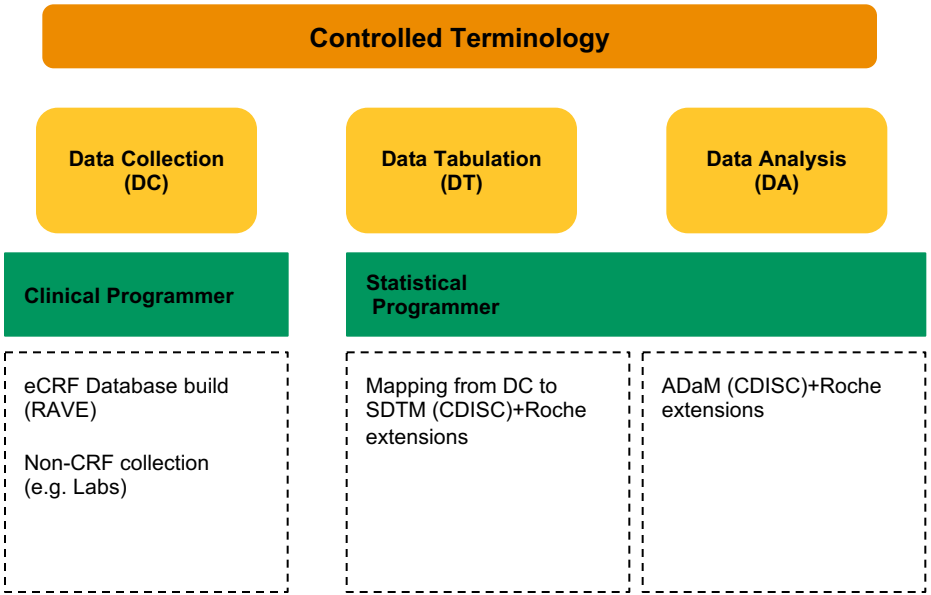
Huw Mason

Outline

- Introduction
- Knowledge Graphs
- What Makes GDSR FAIR?
  - **F**indable
  - **A**ccessible
  - **I**nteroperable
  - **R**eusable
- Summary

Introduction

The Global Data Standards Repository (GDSR) is Roche and Genentech's single source of truth for clinical data standards, storing and making them available to users globally in the most business useful format.



Introduction

Clinical Data Standards
Play a key role in clinical trials

- Key for analysis, sharing and ensuring interoperability

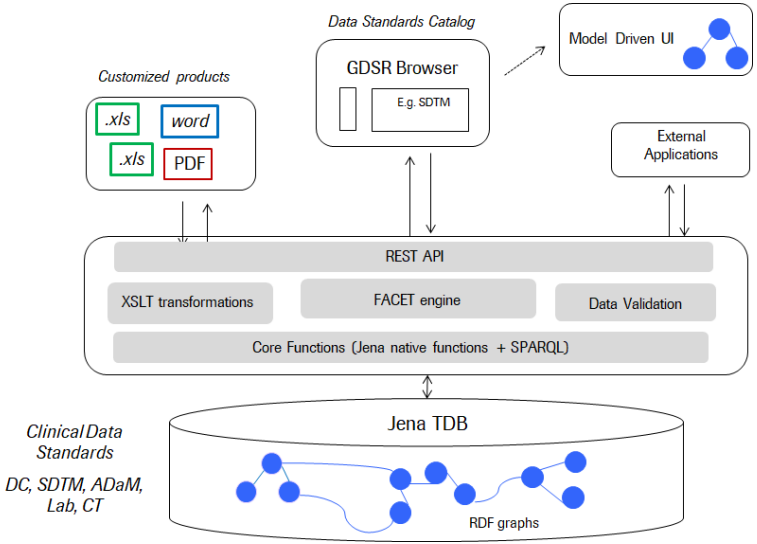
- Roche Global Clinical Data Standards are aligned with CDISC (Clinical Data Interchange Standards Consortium) standards

- From collection to tabulation and analysis

- Metadata about these data standards is managed in the Global Data Standards Repository

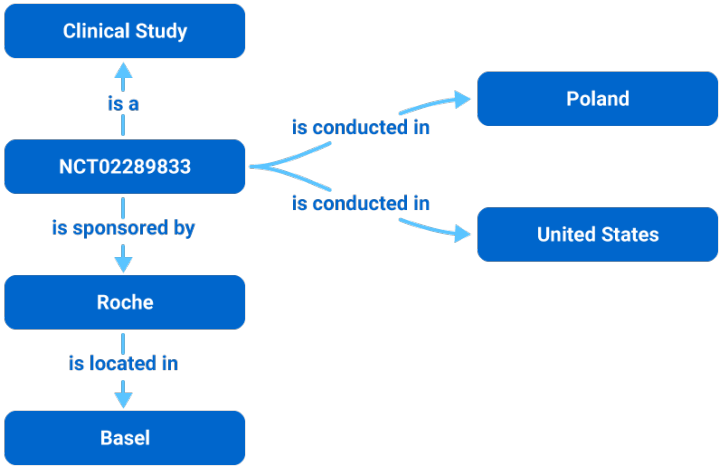
Introduction

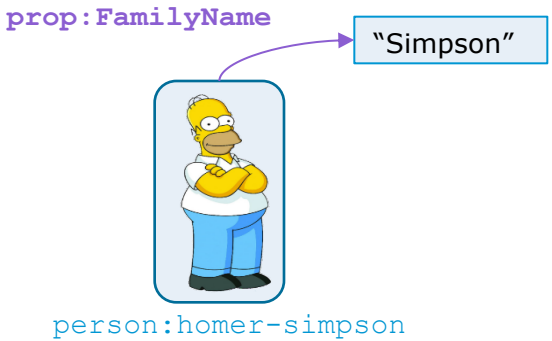
The data standards are stored in semantic graphs and the technology behind the GDSR is based on RDF, ontologies, and linked data principles aligned with Semantic Web (WC3) standards. Having common vocabularies for (meta) data assets ensures common understanding and facilitates integration and sharing for machines.



Knowledge Graphs

- A graph of data with the intention to encode **knowledge**
  - Nodes are **entities** (something in the real world)
  - Labelled **edges** between two nodes capture a relationship between entities
  - **Ontologies** can define possible types of entities, relationships and attributes



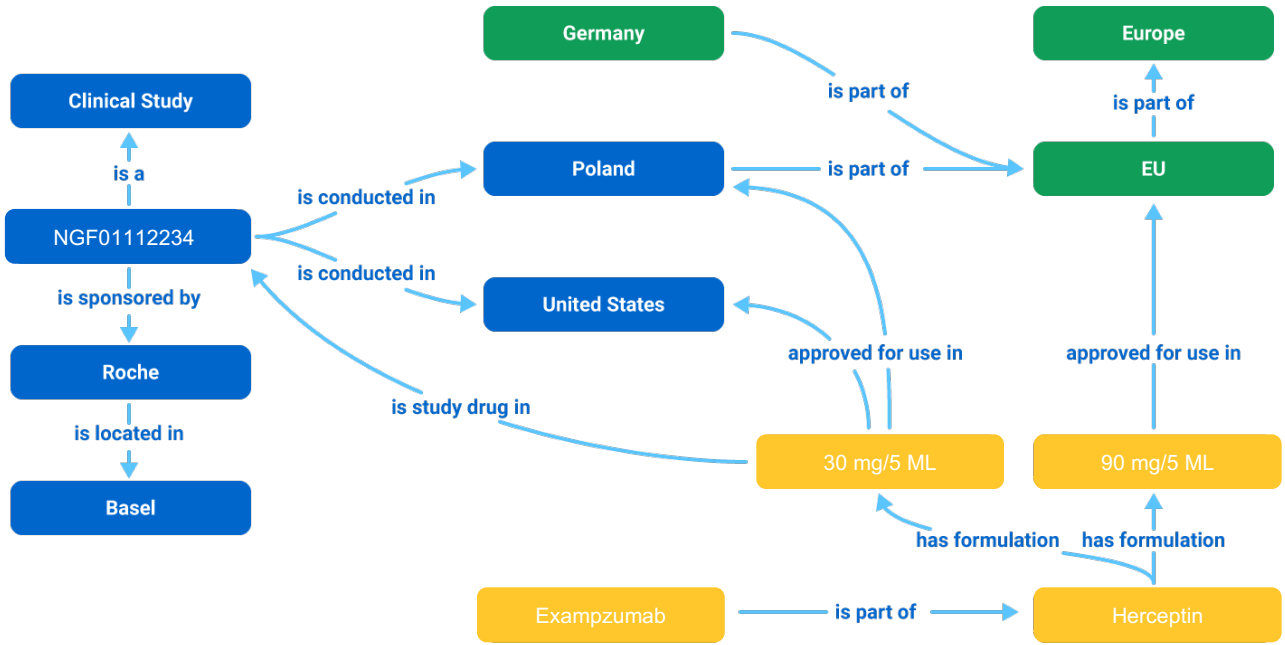


Knowledge Graphs - Why?

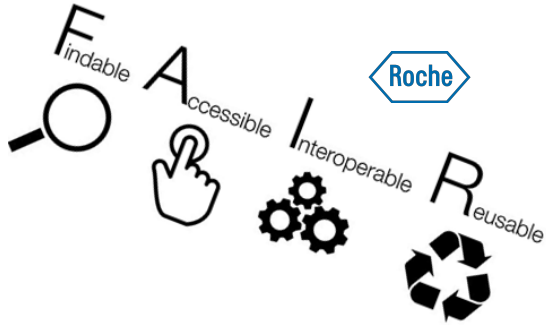
Benefits

- Easy to understand: data is described as human think
- Implemented using **linked data principles**, therefore contributing to **FAIR data**
- **Leverage** ongoing activities using KGs across the organization
- Enables **knowledge discovery** through querying and searching
- Facilitate data integration
    - **Flexible** and **easy to connect decentralized** but related information
    - Possible to **enrich with content** from internal and external sources

Knowledge Graphs



The FAIR principles



Finding data and associated context

Accessing data from different systems

Data **interoperability and sharing** culture – “our” data

Reusing data/identifiers and properly tracking data

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data3 (2016).

The FAIR guiding principles

Designed to enhance reusability of data assets

- To be Findable
  - F1: (meta) data are assigned globally unique and persistent identifiers
  - F2: Data are described with rich metadata
  - F3: Metadata clearly and explicitly include the identifier of the data it describes
  - F4: (meta)data are registered or indexed in a searchable resource
- To be Accessible
  - A1: (meta)data are retrievable by their identifier using a standardised communication protocol.
    - A1.1: The protocol is open, free and universally implementable
    - A1.2: The protocol allows for an authentication and authorisation when required
  - A2: Metadata should be accessible even when the data is no longer available

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data3 (2016).

The FAIR guiding principles

Designed to enhance reusability of data assets

- To be Interoperable
  - I1: (meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation.
  - I2: (meta)data use vocabularies that follow the FAIR principles
  - I3: (meta)data include qualified references to other (meta) data.
- To be Reusable
  - R1: meta(data) are richly described with a plurality of accurate and relevant attributes
    - R1.1: (meta)data are released with a clear and accessible data usage license.
    - R1.2: (meta)data are associated with detailed provenance
    - R1.3: (meta)data meet domain-relevant community standards

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data3 (2016).

What makes GDSR FAIR?

Interoperable data I2: (meta)data use vocabularies that follow the FAIR principles

- Ensure data across sites can be combined easily by machines
- Use of formal data standards and controlled vocabularies

Findable

- Unique and persistent identifier
- Rich Metadata
- Searchable (meta)data

- **GDSR** uses RDF and ontologies to describe the metadata and standard models. Hence, each concept is identified using a URI.

<http://gdsr.roche.com/cdisc/adamig#Table.ADSL>

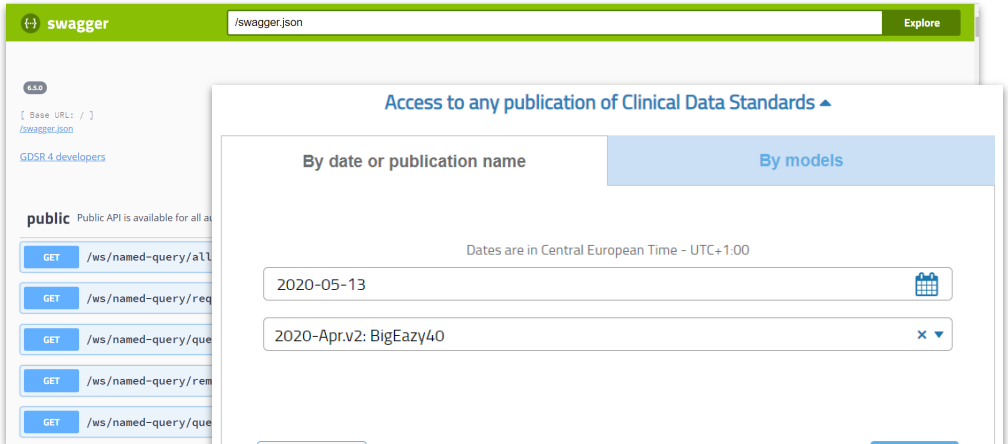
- **Metadata** about the models is provided by a catalog, giving users a summary by publication.
- **Reports** give detail to the components of a model.
- **GDSR** browser supports indexed searching.

Accessible

- Retrievable (meta)data via standard protocols
- (Optional) authentication/ authorization
- Metadata is always accessible

- **GDSR** makes content ACCESSIBLE through a browser and a public REST API, and uses admin roles for authentication.

- **Publication** Versioning allows users to access archives of previous data standards versions.



Interoperable

- meta(data) in a formal and shared language
- external meta(data) are FAIR
- meta(data) points to external (meta)data

- **GDSR** uses RDF/OWL ontologies and RDF for representing metadata.
- **All metadata** and schema in the GDSR are described in ontologies.
- **Standards** stored in the GDSR link to external datasets e.g. CDISC, enabling the exploration of connected vocabularies.

```
<vmms:Model rdf:about="http://gdsr.roche.com/pd-biometrics/models#Model.ADaM">
  <vmms:isDefinedBy rdf:resource="http://gdsr.roche.com/pd-biometrics/models-metadata#Clinical_data_interchange_standards_consortium"/>
  <vmms:isModelType rdf:resource="http://gdsr.roche.com/pd-biometrics/models-metadata#Foundational_model"/>
  <vmms:isUsedFor rdf:resource="http://gdsr.roche.com/pd-biometrics/models-metadata#Data_analysis_clinical_trial"/>
  <skos:altLabel xml:lang="en">ADaM</skos:altLabel>
  <skos:definition xml:lang="en">CDISC foundational model describing the underlying rules for ADaM datasets. New versions are added following a deci
  <skos:note xml:lang="en">When published by CDISC and Roche choose to adopt - should use the CDISC version number</skos:note>
  <skos:prefLabel xml:lang="en">Analysis data model</skos:prefLabel>
</vmms:Model>
```

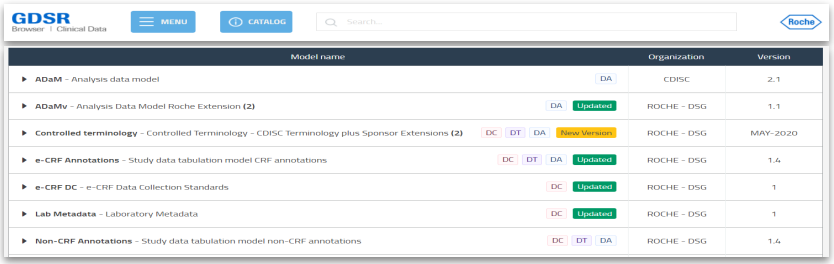
Reusable

- Rich, plural, accurate meta(data)
- Accessible data license
- Keep the provenance of the information

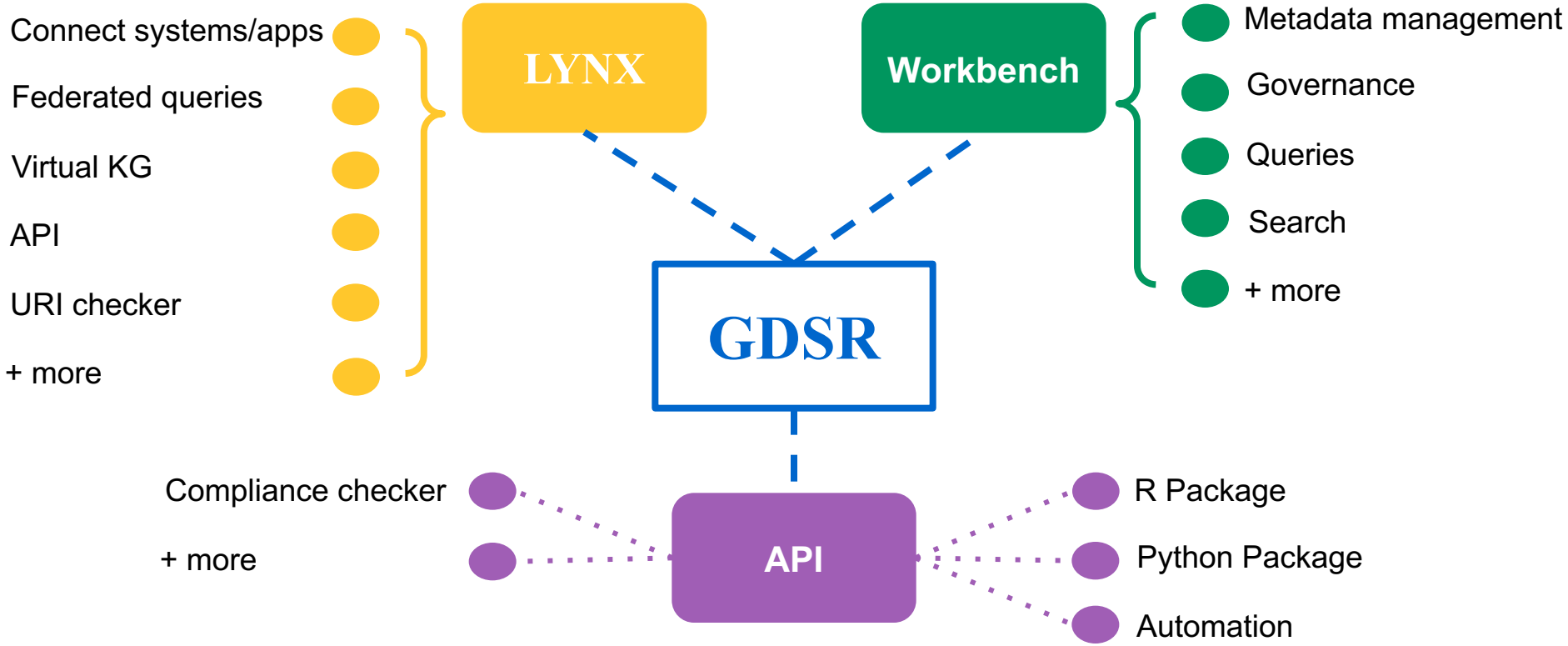
- **The Catalog** contains authorship of Data Standards Models.

- **The Provenance** ontology PROV-O is used to capture information about changes to the models between versions.

- **GDSR** technology follows community trends, it is aligned with Semantic Web (W3) standards and vocabularies, as well as Pharma best practices (CDISC).



Summary


Connect systems/apps
Federated queries
Virtual KG
API
URI checker
+ more
LYNX
Workbench
Metadata management
Governance
Queries
Search
+ more
GDSR
Compliance checker
+ more
API
R Package
Python Package
Automation

Thanks!

Developed by the **RGITSC Custom Apps team** and **DAAV** Information Architects.

Slides, Content, and Materials

Javier D. Fernández, Didier Clement, Nelia Lasierra, Huw Mason, Adam Kozak, Michal Pietrusinski, Kamil Tatara, Matt Groening

DAAV - Information Architects

Selena Baset, Didier Clement, Javier D. Fernández, Nelia Lasierra, Huw Mason, Elvira Mitraka, Ivan Robinson

RGITSC team

Adam Kozak,Michal Pietrusinski, Michal Sadowski, Jakub Urbanowicz, Jakub Wach, Michal Kalinowski, Celso Pereira, Bartosz Grab, Chrystian Klingenberg, Ewa Patrzykat, Sebastian Choina, Jacek Tominski, Jacek Caban, Alicja Bialek, David Rycerz, Kamil Bojarski, Michal Kalinowski, Katarzyna Ulmer, Kamil Tatara, Adam Goralczyk, Anna Grzymala, Dariusz Stepien, Jakub Pazdyga, Krzysztof Nielepkowicz, Szymon Tyka, Dawid Rycerz, Adam Lempart + Operations Team

Questions

Doing now what patients need next

Backup

nature > scientific data > comment > article

SCIENTIFIC DATA

Comment | Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, [...] Barend Mons

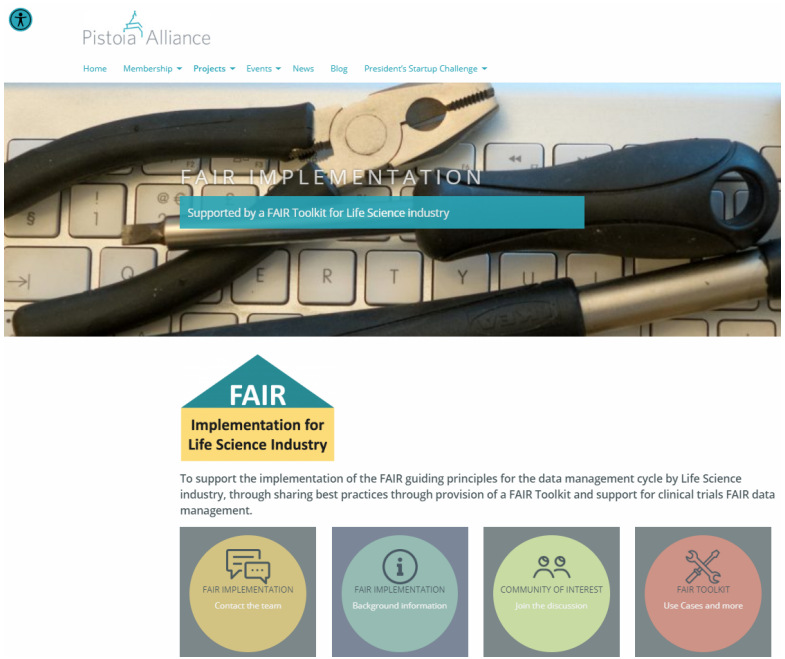
Scientific Data **3**, Article number: 160018 (2016) | Cite this article

84k Accesses | **1158** Citations | **1459** Altmetric | Metrics

An Addendum to this article was published on 19 March 2019

Abstract

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that



HCLS community profile --> fairsharing.org

fairmetrics.org