Immunological Data Standards and Repositories

R. Burke Squires

Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

This is a human-readable summary of (and not a substitute for) the license. Disclaimer.

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

Please Note: parts of this work have been borrowed from other individuals and may contain a different license.

Objectives

- Enable you to name one or more immunologically relevant data standards
- Enable you to search for immunologically relevant data repositories
- Enable you to take a step in becoming a reproducible scientist

Outline

- Immunological Data
  - Standards
  - Repositories

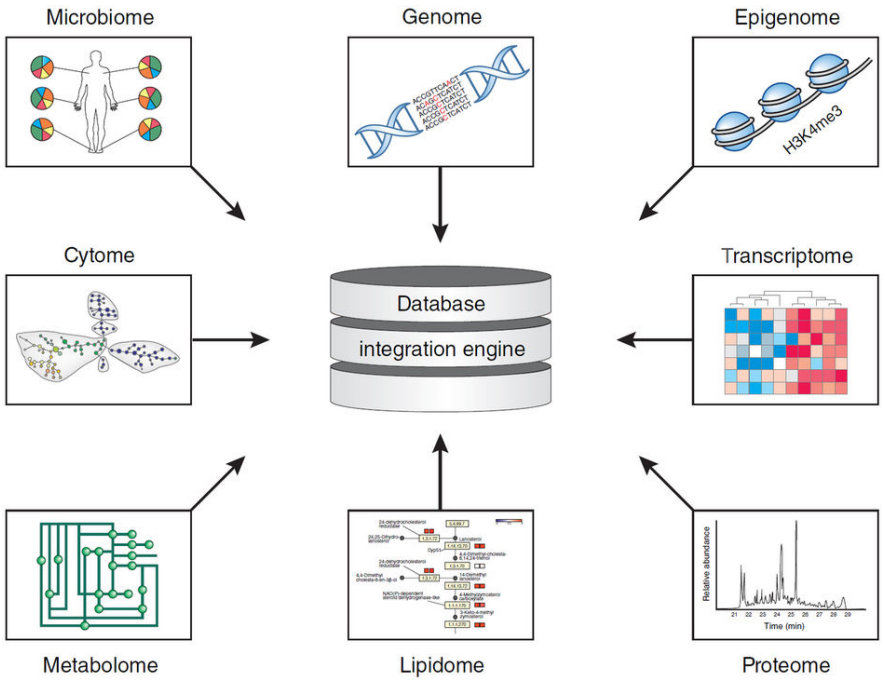


Image: Schultze, J. L. Teaching 'big data' analysis to young immunologists. *Nature Immunology* **16,** 902–905 (2015).

Immunological Data

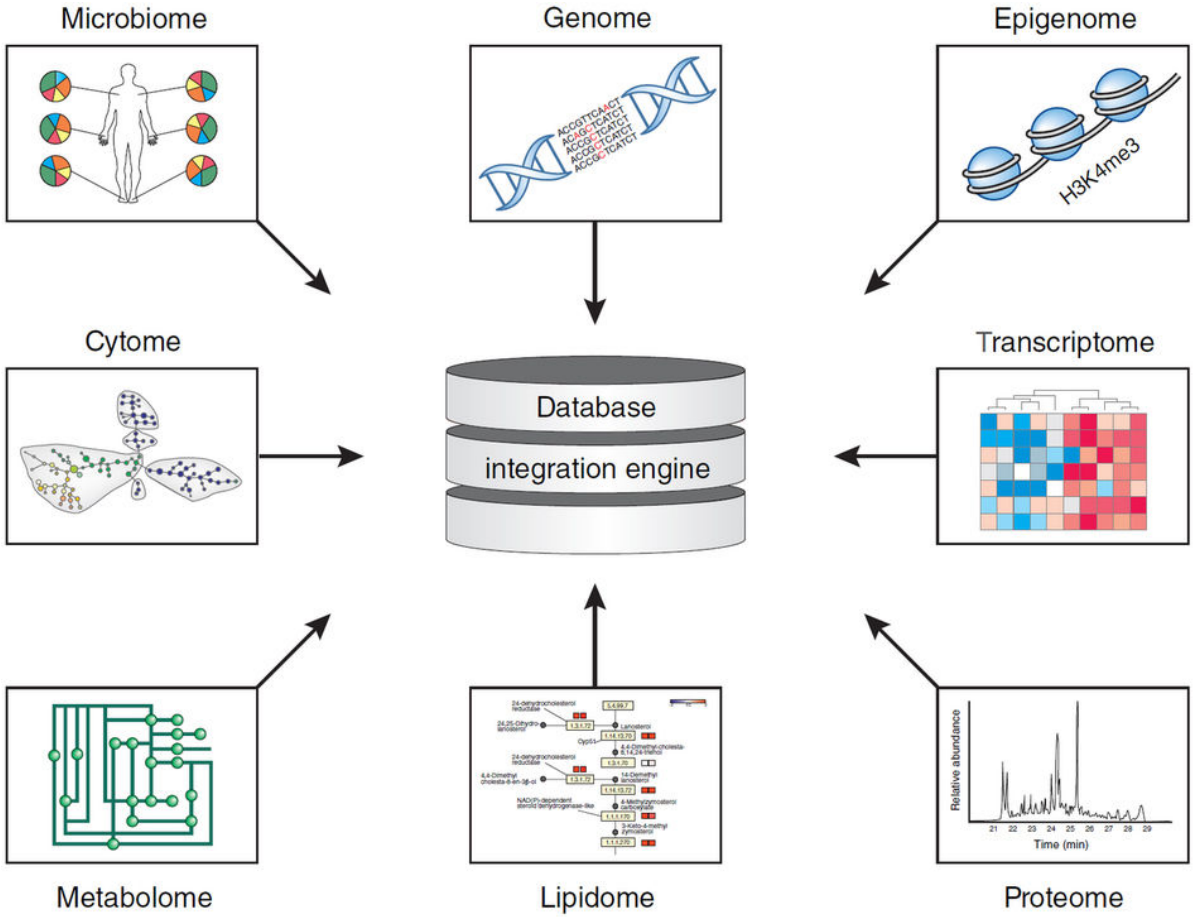


Image: Schultze, J. L. Teaching 'big data' analysis to young immunologists. *Nature Immunology* **16,** 902–905 (2015).

Data Standards

Data Standards

- What do we mean by standards
- Why have them?
- What is meant by minimal information?
- Where can you find data standards?
- What are some immunologically relevant data standards?

Data Standards:
Standards – a Definition

- Agreed-upon conventions for doing ‘something’, established by community consensus or an authority
  - e.g. managing a process or delivering a service

Data Standards: Interoperability Standards

- Nuts and Bolts
- Standardization initiated in 1864 by William Sellers
- Only widely adopted a century later!

Data Standards: Interoperability Standards

- Agreed-upon specifications, guidelines or criteria designed to ensure data and any other digital object (such as code, algorithms, workflows, models, software, or journal articles) are FAIR

- FAIR
  - Findable
  - Accessible
  - Interoperable
  - Reusable

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

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 focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes 'good data management' is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessibility, Interoperability, and Reusability—that serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing. Importantly, it is our intent that the principles apply not only to 'data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects[2]—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other's data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly

Correspondence and requests for materials should be addressed to B.M. (email: barend.mons@dtls.nl).
#A full list of authors and their affiliations appears at the end of the paper.

Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3,** 160018 (2016).

Data Standards:
Interoperability Standards – Fundamentals

- Essential for discovery of data and any other digital object, but also citation and credit

- Enable the operational processes, underlying their (re)use
  - such as exchange, aggregation, integration, comparison

- Optimal interoperability is reached when all processes are automated for both human and machine, this requires
  - metadata: or descriptors for the digital objects to understand what it is, where to find it, how to access it etc.
  - identifiers: unique, resolvable and versionable identifiers are essential elements of the digital word

Data Standards:
Interoperability Standards – Realizing Potential

- Identifiers and metadata must be implemented by technical experts in tools, registries, catalogues, databases, services
  - to find, store, manage (e.g., mint, track provenance, version) and aggregate (e.g., interlink and map etc.) these digital objects.

- Implementations are essential to make standards 'invisible' to users, such as researchers, who often have little or no familiarity with them

Data Standards:
Metadata Standards – Fundamentals

- Metadata - data about the data!

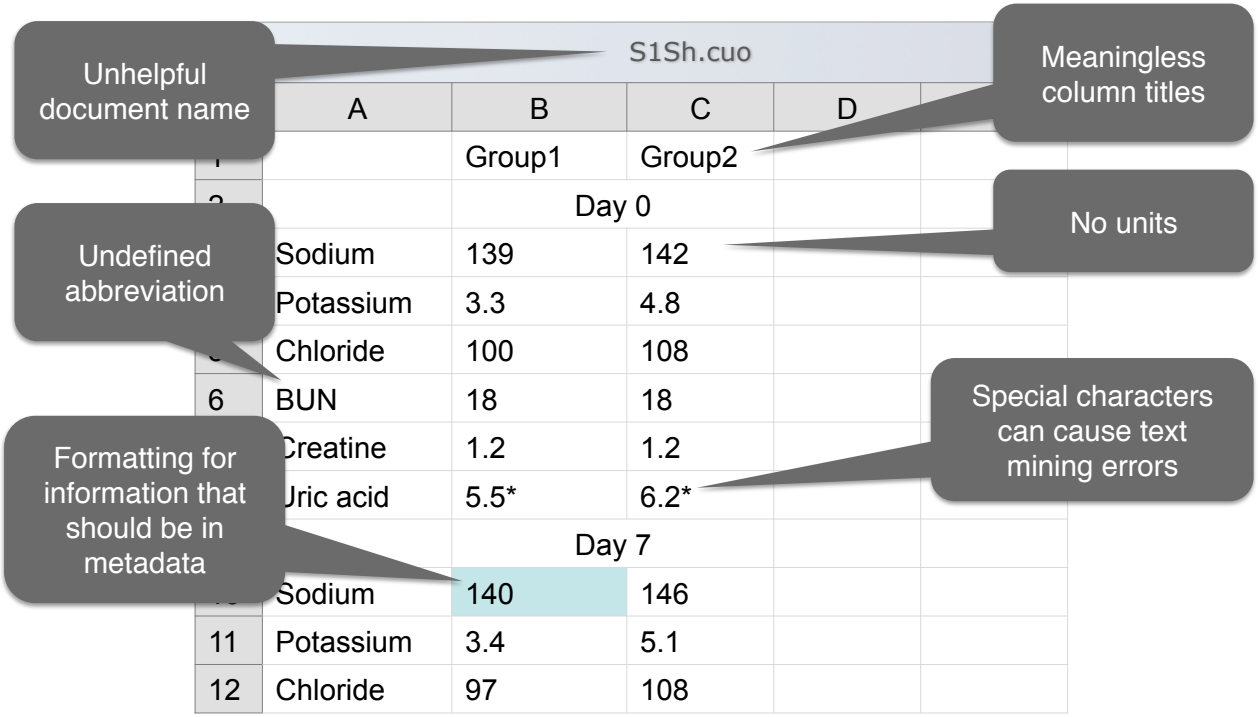
- Descriptors for a digital object that help to understand what it is, where to find it, how to access it etc.

- The type of metadata depends also on the digital object

- The depth and breadth of metadata varies according to their purpose (discovery, citation, credit)
  - e.g. reproducibility requires richer metadata

Data Standards:
Metadata Standards – Fundamentals

Sharing starts with good metadata…but this not!



Data Standards:
Metadata Standards – Fundamentals

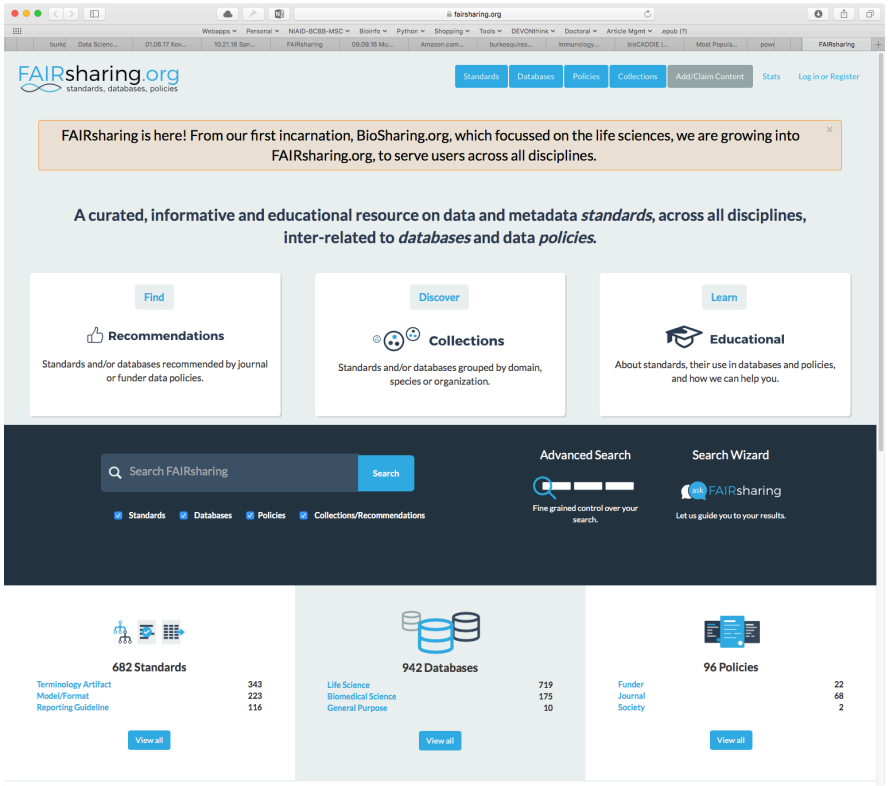
….this is much clearer!

Table_S1_Shanghai_blood.xls

	A	B	C	D	E	F
1	Parameter	Day	Control	Treated	Units	P
2	Sodium	0	139	142	mEq/l	0.82
3	Sodium	7	140	146	mEq/l	0.70
4	Sodium	14	140	158	mEq/l	0.03
5	Sodium	21	143	160	mEq/l	0.02
6	Potassium	0	3.3	4.8	mEq/l	0.06
7	Potassium	7	3.4	5.1	mEq/l	0.07
8	Potassium	14	3.7	4.7	mEq/l	0.10
9	Potassium	21	3.1	3.6	mEq/l	0.52
10	Chloride	0	100	108	mEq/l	0.56
11	Chloride	7	97	108	mEq/l	0.68
12	Chloride	14	101	106	mEq/l	0.79

Data Standards: Searching for Standards

- Where do we find them?
- Biosharing.org is now FAIRsharing.org
- FAIRsharing.org
  - Standards
  - Databases
  - Policies
  - Collections
  - Add…



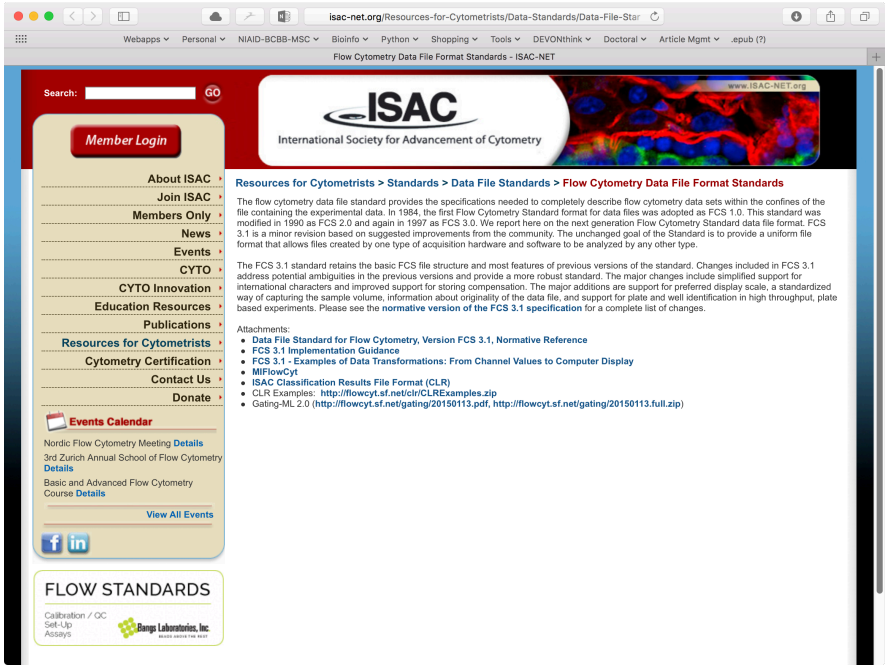
Demo

What Are Some Immunologically Relevant Data Standards?

- Flow Cytometry Data File Format Standards (FCS 3.1)
- ISAC Classification Results File Format (CLR)
- ISAC standard for representing gating descriptions in flow cytometry (Gating-ML)
- Minimum Information for Biological and Biomedical Investigations (MIBBI)
  - Minimum Information About a Bioinformatics investigation (MIABI)
  - Minimum Information about a Flow Cytometry Experiment (MIFlowCyt)
- LINCS Extended Metadata Standard
  - Antibody reagents
  - Cell lines
  - Differentiated cells
  - Embryonic stem cells
  - iPSCs
  - Nucleic acid reagents
  - Other reagents
  - Primary cells
  - Proteins
  - Small molecules

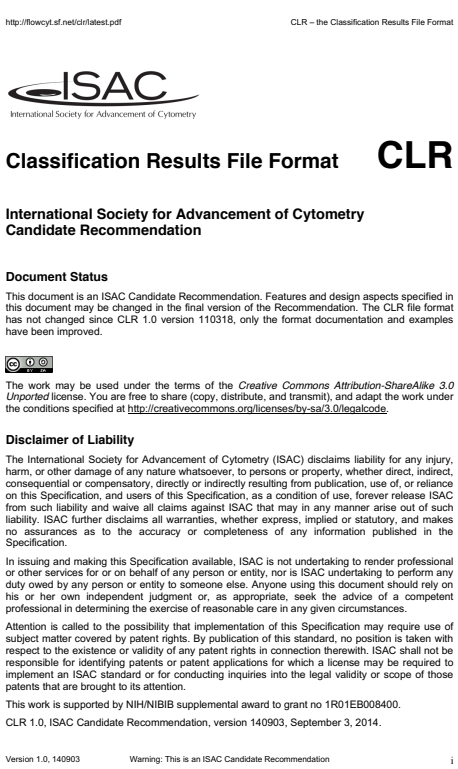
Flow Cytometry Data File Format Standards (FCS)

- The flow cytometry data file standard provides the specifications needed to completely describe flow cytometry data sets within the confines of the file containing the experimental data.
- Goal of the Standard is to provide a uniform file format that allows files created by one type of acquisition hardware and software to be analyzed by any other type.
- FCS Is implemented in
  - FlowRepository
  - The Immunology Database and Analysis Portal - OpenImmport



ISAC Classification Results File Format (CLR)

- CLR has been developed to exchange the results of manual gating and algorithmic classification approaches in a standard way in order to be able to report and process the classification.
- Although it was originally designed for the field of flow cytometry, it is applicable in any domain that needs to capture either soft or unambiguous classifications of virtually any kinds of objects.



ISAC
International Society for Advancement of Cytometry

Classification Results File Format **CLR**

International Society for Advancement of Cytometry
Candidate Recommendation

Document Status

This document is an ISAC Candidate Recommendation. Features and design aspects specified in this document may be changed in the final version of the Recommendation. The CLR file format has not changed since CLR 1.0 version 110318, only the format documentation and examples have been improved.

Disclaimer of Liability

The International Society for Advancement of Cytometry (ISAC) disclaims liability for any injury, harm, or other damage of any nature whatsoever, to persons or property, whether direct, indirect, consequential or compensatory, directly or indirectly resulting from publication, use of, or reliance on this Specification, and users of this Specification, as a condition of use, forever release ISAC from such liability and waive all claims against ISAC that may in any manner arise out of such liability. ISAC further disclaims all warranties, whether express, implied or statutory, and makes no assurances as to the accuracy or completeness of any information published in the Specification.

In issuing and making this Specification available, ISAC is not undertaking to render professional or other services for or on behalf of any person or entity, nor is ISAC undertaking to perform any duty owed by any person or entity to someone else. Anyone using this document should rely on his or her own independent judgment or, as appropriate, seek the advice of a competent professional in determining the exercise of reasonable care in any given circumstances.

Attention is called to the possibility that implementation of this Specification may require use of subject matter covered by patent rights. By publication of this standard, no position is taken with respect to the existence or validity of any patent rights in connection therewith. ISAC shall not be responsible for identifying patents or patent applications for which a license may be required to implement an ISAC standard or for conducting inquiries into the legal validity or scope of those patents that are brought to its attention.

This work is supported by NIH/NIBIB supplemental award to grant no 1R01EB008400.

CLR 1.0, ISAC Candidate Recommendation, version 140903, September 3, 2014.

http://isac-net.org/PDFS/65/659585c7-9ef4-4152-829c-859aad25aafb.pdf

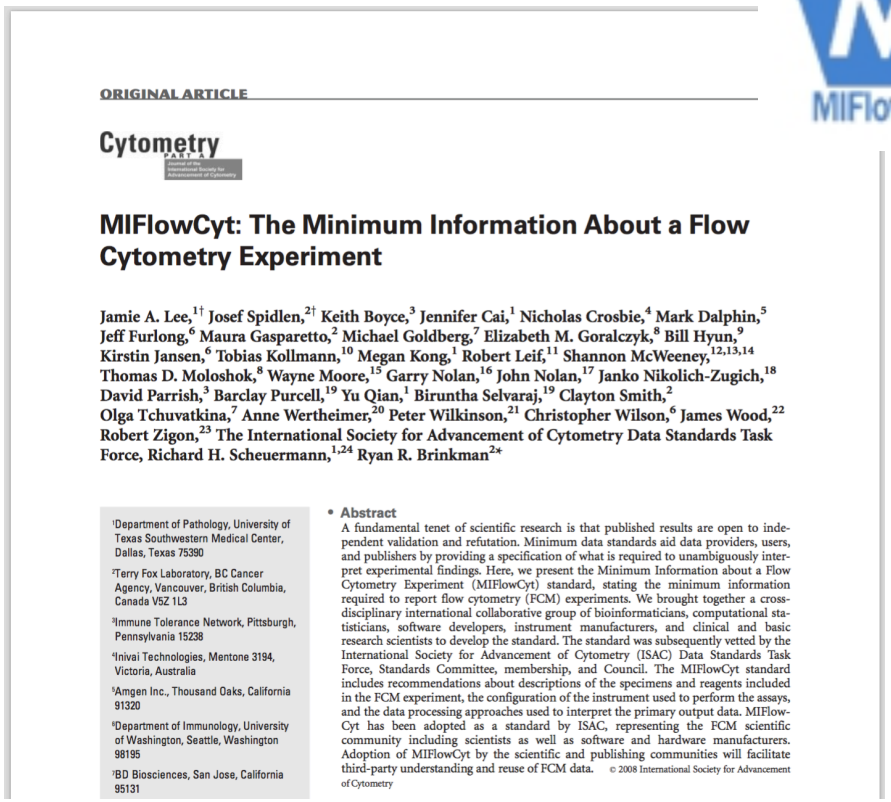
NIH Library of Integrated Network-Based Cellular Signatures (LINCS)

- The Library of Integrated Network-Based Cellular Signatures (LINCS) Program aims to create a network-based understanding of biology by cataloging changes in gene expression and other cellular processes that occur when cells are exposed to a variety of perturbing agents.

- Data portal
  - http://lincsportal.ccs.miami.edu/dcic-portal/
- Cells
  - http://lincsportal.ccs.miami.edu/cells/

Minimum Information About a Flow Cytometry Experiment (MIFlowCyt)

- Minimum Information about a Flow Cytometry Experiment (MIFlowCyt)
  - Outlines the minimum information required to report about flow cytometry experiments
  - Represents the community consensus
    - 33 coauthors from 19 institutions
    - ISAC Recommendation
  - Required/recommended by Cytometry A and Nature



ORIGINAL ARTICLE

Cytometry PART A

Journal of the International Society for Advancement of Cytometry

MIFlowCyt: The Minimum Information About a Flow Cytometry Experiment

Jamie A. Lee,[1†] Josef Spidlen,[2†] Keith Boyce,[3] Jennifer Cai,[1] Nicholas Crosbie,[4] Mark Dalphin,[5] Jeff Furlong,[6] Maura Gasparetto,[2] Michael Goldberg,[7] Elizabeth M. Goralczyk,[8] Bill Hyun,[9] Kirstin Jansen,[6] Tobias Kollmann,[10] Megan Kong,[1] Robert Leif,[11] Shannon McWeeney,[12,13,14] Thomas D. Moloshok,[8] Wayne Moore,[15] Garry Nolan,[16] John Nolan,[17] Janko Nikolich-Zugich,[18] David Parrish,[3] Barclay Purcell,[19] Yu Qian,[1] Biruntha Selvaraj,[19] Clayton Smith,[2] Olga Tchuvatkina,[7] Anne Wertheimer,[20] Peter Wilkinson,[21] Christopher Wilson,[6] James Wood,[22] Robert Zigon,[23] The International Society for Advancement of Cytometry Data Standards Task Force, Richard H. Scheuermann,[1,24] Ryan R. Brinkman[2*]

[1]Department of Pathology, University of Texas Southwestern Medical Center, Dallas, Texas 75390

[2]Terry Fox Laboratory, BC Cancer Agency, Vancouver, British Columbia, Canada V5Z 1L3

[3]Immune Tolerance Network, Pittsburgh, Pennsylvania 15238

[4]Inivai Technologies, Mentone 3194, Victoria, Australia

[5]Amgen Inc., Thousand Oaks, California 91320

[6]Department of Immunology, University of Washington, Seattle, Washington 98195

[7]BD Biosciences, San Jose, California 95131

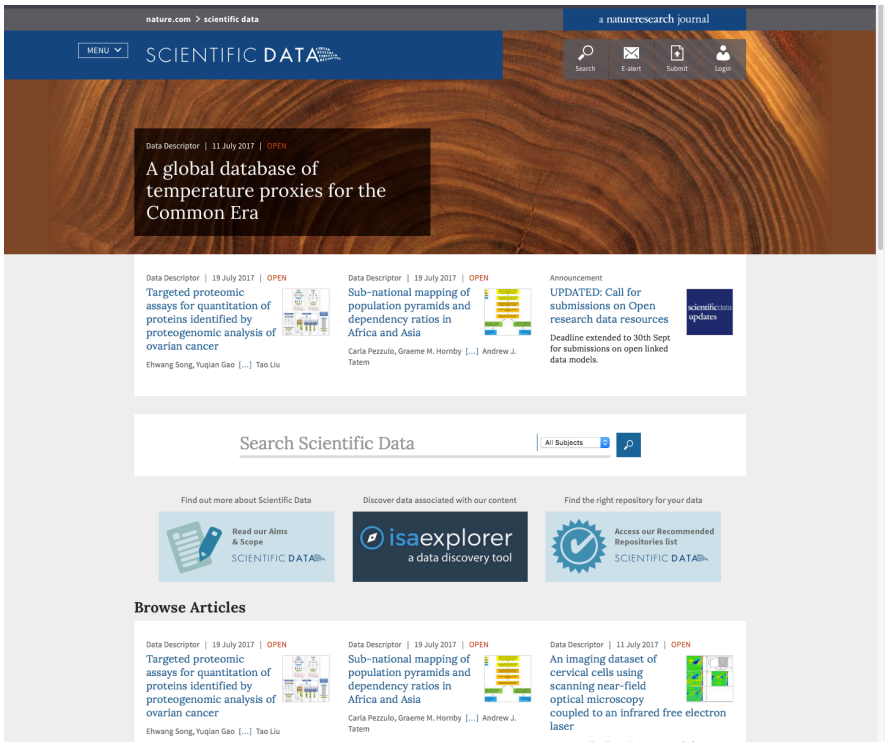
• Abstract

A fundamental tenet of scientific research is that published results are open to independent validation and refutation. Minimum data standards aid data providers, users, and publishers by providing a specification of what is required to unambiguously interpret experimental findings. Here, we present the Minimum Information about a Flow Cytometry Experiment (MIFlowCyt) standard, stating the minimum information required to report flow cytometry (FCM) experiments. We brought together a cross-disciplinary international collaborative group of bioinformaticians, computational statisticians, software developers, instrument manufacturers, and clinical and basic research scientists to develop the standard. The standard was subsequently vetted by the International Society for Advancement of Cytometry (ISAC) Data Standards Task Force, Standards Committee, membership, and Council. The MIFlowCyt standard includes recommendations about descriptions of the specimens and reagents included in the FCM experiment, the configuration of the instrument used to perform the assays, and the data processing approaches used to interpret the primary output data. MIFlowCyt has been adopted as a standard by ISAC, representing the FCM scientific community including scientists as well as software and hardware manufacturers. Adoption of MIFlowCyt by the scientific and publishing communities will facilitate third-party understanding and reuse of FCM data.

Data as a Primary Resource

Data as a Primary Resource

- It gets it's own journal – from Nature Publishing Group!
- *Scientific Data*



Not Everyone is Happy With Data Reuse: “Research Parasites”

- On January 21, 2016, the Editors of the New England Journal of Medicine, Dan L. Longo, M.D., and Jeffrey M. Drazen, M.D., published an editorial that characterized [some] scientists who re-analyzed published data sets as "research parasites"

The NEW ENGLAND JOURNAL of MEDICINE

EDITORIALS

Data Sharing

Dan L. Longo, M.D., and Jeffrey M. Drazen, M.D.

The aerial view of the concept of data sharing is beautiful. What could be better than having high-quality information carefully reexamined for the possibility that new nuggets of useful data are lying there, previously unseen? The potential for leveraging existing results for even more benefit pays appropriate increased tribute to the patients who put themselves at risk to generate the data. The moral imperative to honor their collective sacrifice is the trump card that takes this trick.

However, many of us who have actually conducted clinical research, managed clinical studies and data collection and analysis, and curated data sets have concerns about the details. The first concern is that someone not involved in the generation and collection of the data may not understand the choices made in defining the parameters. Special problems arise if data are to be combined from independent studies and considered comparable. How heterogeneous were the study populations? Were the eligibility criteria the same? Can it be assumed that the differences in study populations, data collection and analysis, and treatments, both protocol-specified and unspecified, can be ignored?

A second concern held by some is that a new class of research person will emerge — people who had nothing to do with the design and execution of the study but use another group's data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited. There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as “research parasites.”

This issue of the *Journal* offers a product of data sharing that is exactly the opposite. The new investigators arrived on the scene with their own ideas and worked symbiotically, rather than parasitically, with the investigators holding the data, moving the field forward in a way that neither group could have done on its own. In this case, Dalerba and colleagues[1] had a hypothesis that colon cancers arising from more primitive colon epithelial precursors might be more aggressive tumors at greater risk of relapse and might be more likely to benefit from adjuvant treatment. They found a gene whose expression appeared to correlate with the expression of genes that characterize more mature colon cancers on gene-expression arrays and whose product was reliably measurable in resected colon cancer specimens by immunohistochemistry. To assess the clinical value of this potential biomarker, they needed a sufficiently large group of patients whose archived tissues could be used to assess biomarker expression and who had been treated in relatively homogeneous way.

They proposed a collaboration with the National Surgical Adjuvant Breast and Bowel Project (NSABP) cooperative group, a research consortium funded by the National Cancer Institute that has conducted seminal research in the treatment of breast and bowel cancer for the past 50 years. The NSABP provided access to tissue and to clinical trial results on an individual patient basis. This symbiotic collaboration found that a small proportion (4%) of colon cancers did not express the biomarker and that the survival of patients with those tumors was poorer than that of patients whose tumors expressed the biomarker. Furthermore, when the effect of adjuvant chemotherapy was assessed, nearly all

Longo, D. L. & Drazen, J. M. Data Sharing. *N Engl J Med* **374,** 276–277 (2016).

Immunological Repositories

Repositories

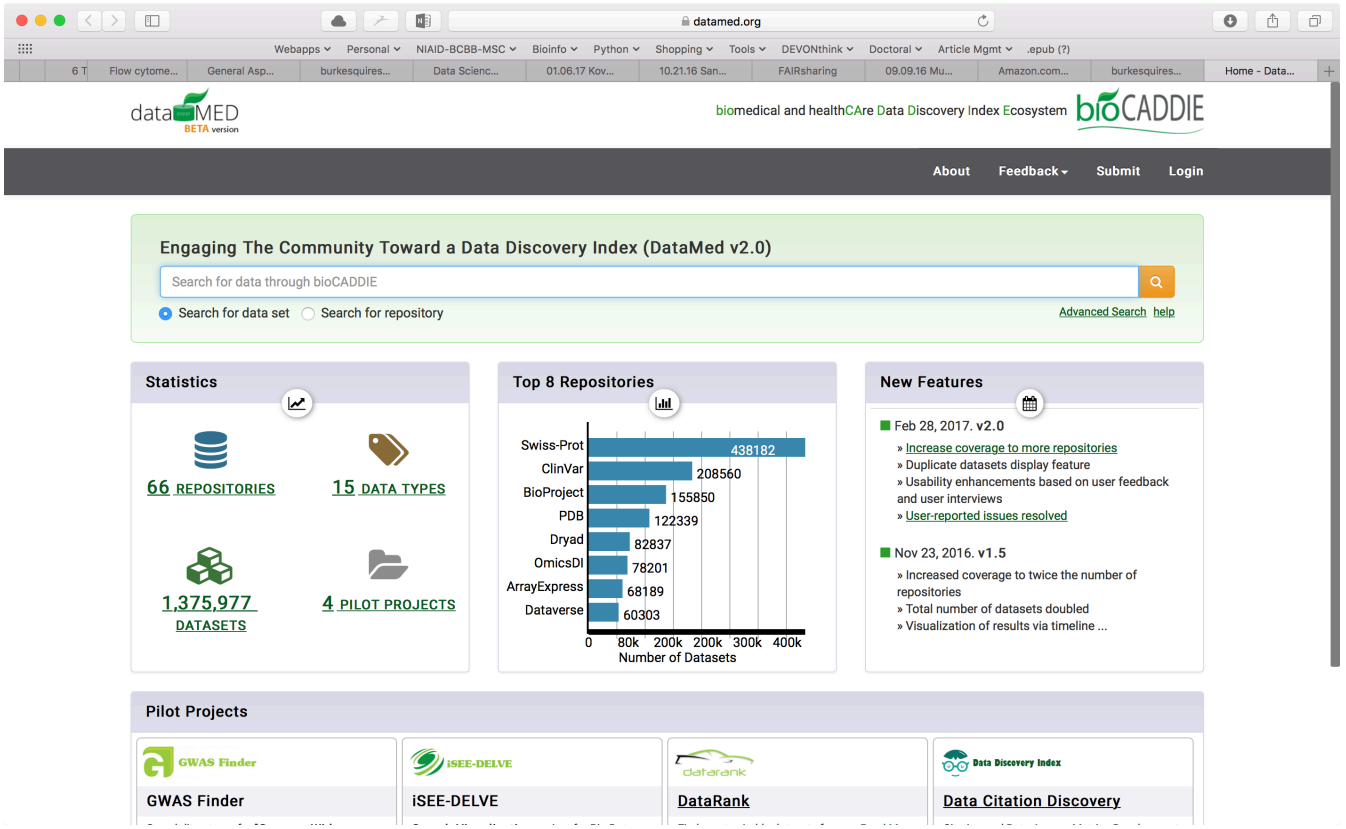
Data (meta)Repositories

- DataMed
  - Data repository search engine
- Nucleic Acid Research (NAR)
  - Database Issues
  - Website Issues
- Review Some Data Repositories

Tool (meta)Repositories

- OMICSTools
  - Immunological tools
- LabWorm
  - Immunological tools

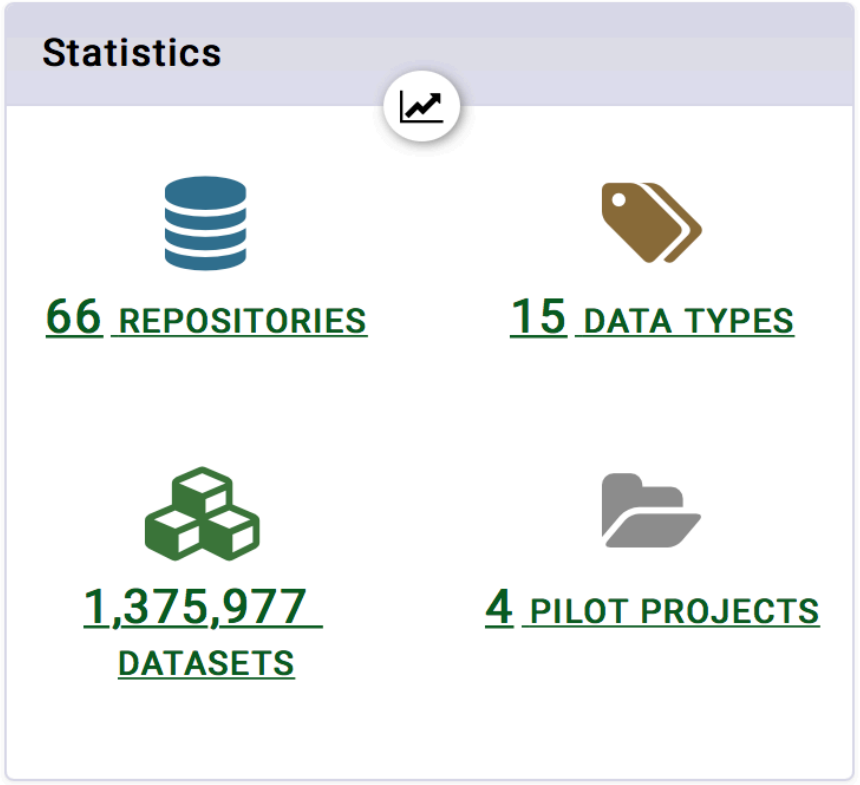
DataMed



Ohno-Machado, L. *et al.* Finding useful data across multiple biomedical data repositories using DataMed. *Nat Genet* **49,** 816–819 (2017).

Repositories: DataMed

- “Immunology”
  - Repositories
    - Dryad (98)
    - BioProject (71)
    - PDB (67)
    - ArrayExpress (55)
    - OmicsDI (53)
    - ImmPort (45)
    - RGD (16)
    - dbGaP (12)
    - GEMMA (4)
    - ProteomeXchange (1)
    - VectorBase (1)
    - Zenodo (1)

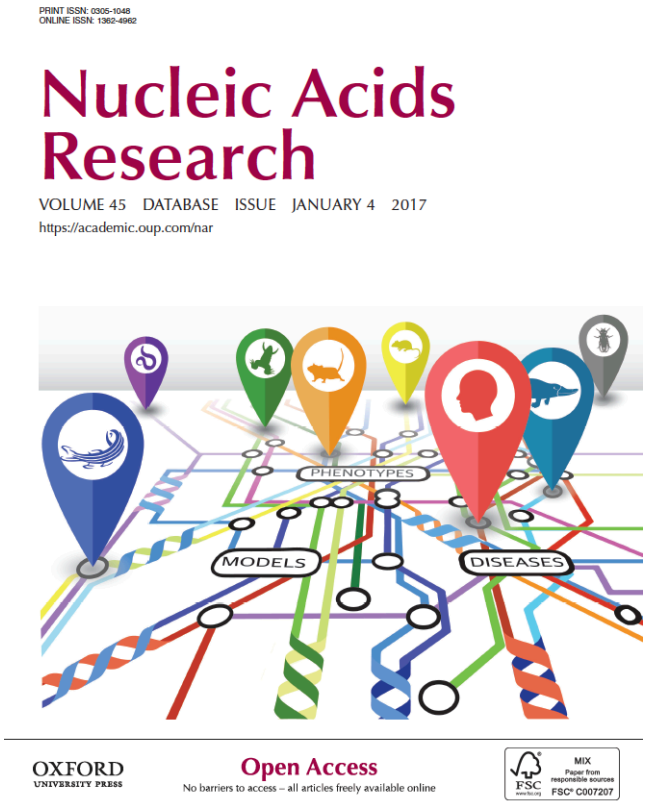


Ohno-Machado, L. *et al.* Finding useful data across multiple biomedical data repositories using DataMed. *Nat Genet* **49,** 816–819 (2017).

Demo

Nucleic Acid Research (NAR) Database Issue

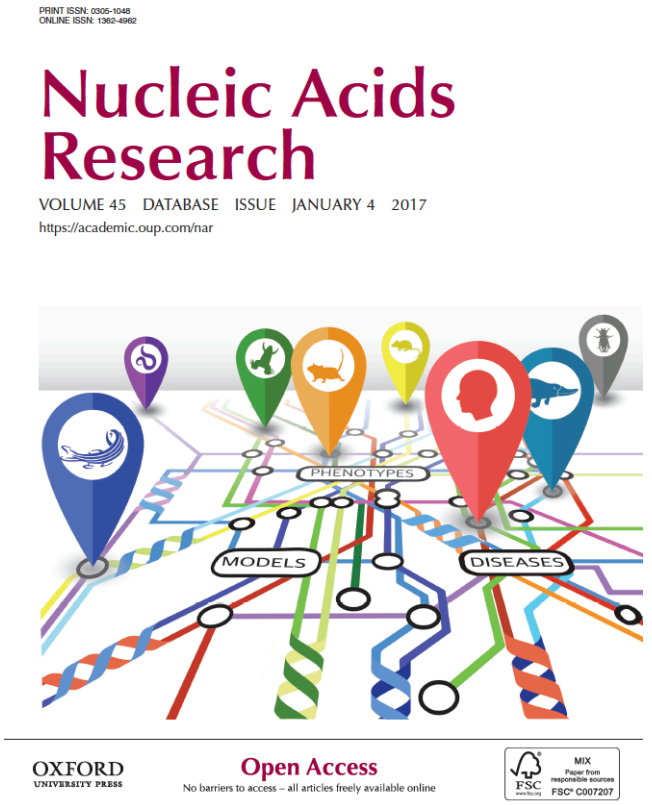
- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- Human and other Vertebrate Genomes
- Human Genes and Diseases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases
- Organelle databases
- Plant databases
- Immunological databases
  - http://www.oxfordjournals.org/nar/database/c
- Cell biology
  - http://www.oxfordjournals.org/nar/database/c



http://www.oxfordjournals.org/nar/database/

Nucleic Acid Research (NAR) Database Issues: Immunological Databases

- AAgAtlas
- ALPSbase
- AntigenDB
- AntiJen
- BCIpep
- bNAber
- dbMHC
- DIGIT
- Epitome
- GPX-Macrophage Expression Atlas
- HaptenDB
- HPTAA
- IEDB
- IEDB-3D
- IL2Rgbase
- IMGT
- IMGT/GENE-DB
- IMGT/HLA
- IMGT/LIGM-DB
- IMGT/mAb-DB
- ImmuNet
- InnateDB
- Interferon Stimulated Gene Database
- IPD - Immuno Polymorphism Database
- IPD-ESTDAB
- IPD-HPA - Human Platelet Antigens
- IPD-KIR - Killer-cell Immunoglobulin-like Receptors
- IPD-MHC
- MHCBN
- MHCPEP
- MPID-T2
- MUGEN Mouse Database
- Protegen
- SAbDab
- SuperHapten
- VBASE2

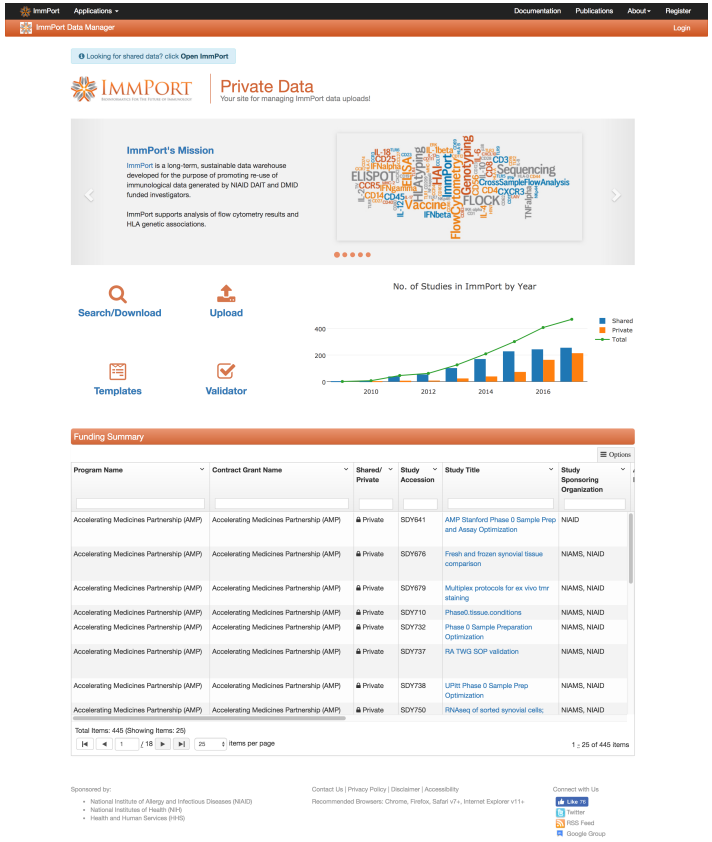


http://www.oxfordjournals.org/nar/database/cat/14

Data Repositories

Immunology Database and Analysis Portal (ImmPort)

- "The goals of the ImmPort project are to:
  - Provide an open access platform for research data sharing
  - Create an integrated environment that broadens the usefulness of scientific data and advances hypothesis-driven and hypothesis-generating research
  - Accelerate scientific discovery while extending the value of scientific data in all areas of immunological research
  - Promote rapid availability of important findings, making new discoveries available to the research community for further analysis and interpretation
  - Provide analysis tools to advance research in basic and clinical immunology"
- Private data and pre-release data are stored in private workspaces of investigators at the ImmPort site located at NIAID, http://immport.niaid.nih.gov.

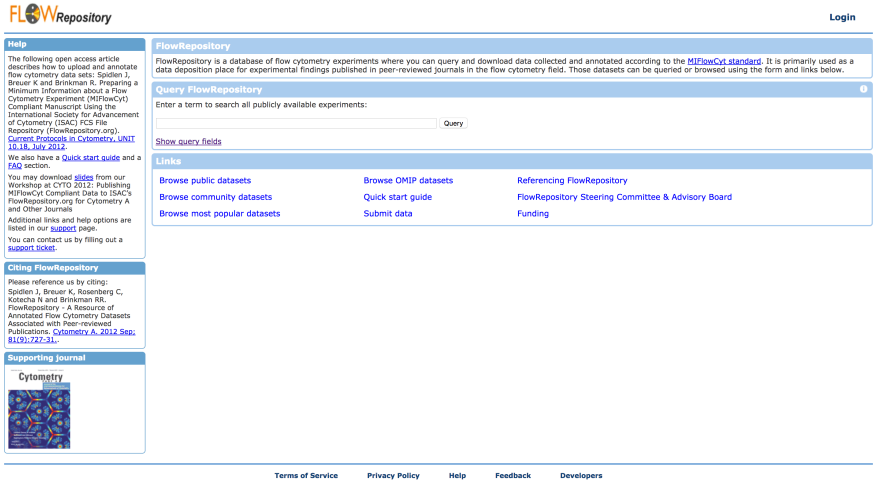


http://immport.niaid.nih.gov

Demo

FlowRepository

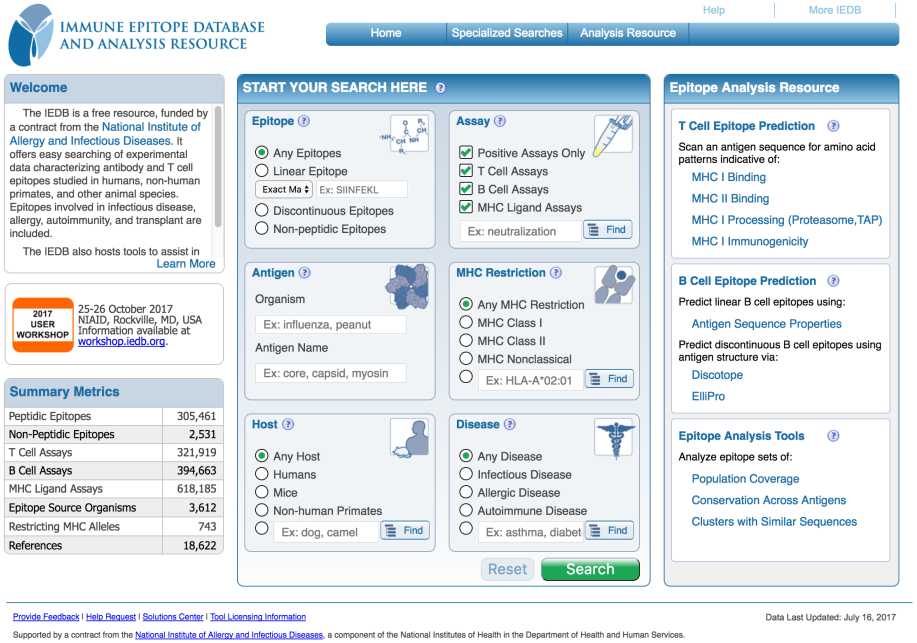
- A database of flow cytometry experiments where you can query and download data collected and annotated according to the MIFlowCyt standard.



Demo

Immune Epitope Database (IEDB)

- Free resource that offers easy searching of experimental data characterizing antibody and T cell epitopes studied in humans, non-human primates, and other animal species.
- Epitopes involved in infectious disease, allergy, autoimmunity, and transplant are included.
- Hosts tools to assist in the prediction and analysis of B cell and T cell epitopes.
- A free resource, funded by a contract from the National Institute of Allergy and Infectious Diseases

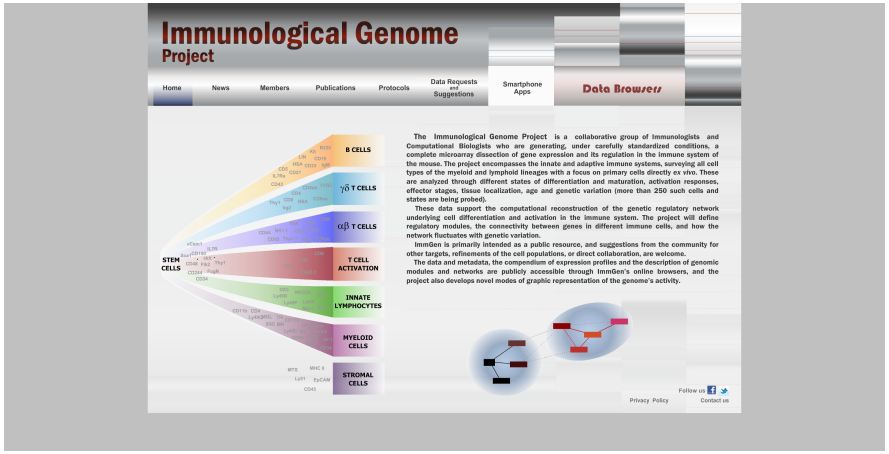


http://www.iedb.org/

Demo

Immunological Genome Project (ImmGen)

- Goal of the project is to computationally reconstruct the gene regulatory network in immune cells
- A gene-expression database for all characterized immune cells in the mouse
- Compendium of microarray data currently include over 250 immunologically relevant cell types, from all lymphoid organs and other tissues which are monitored by immune cells



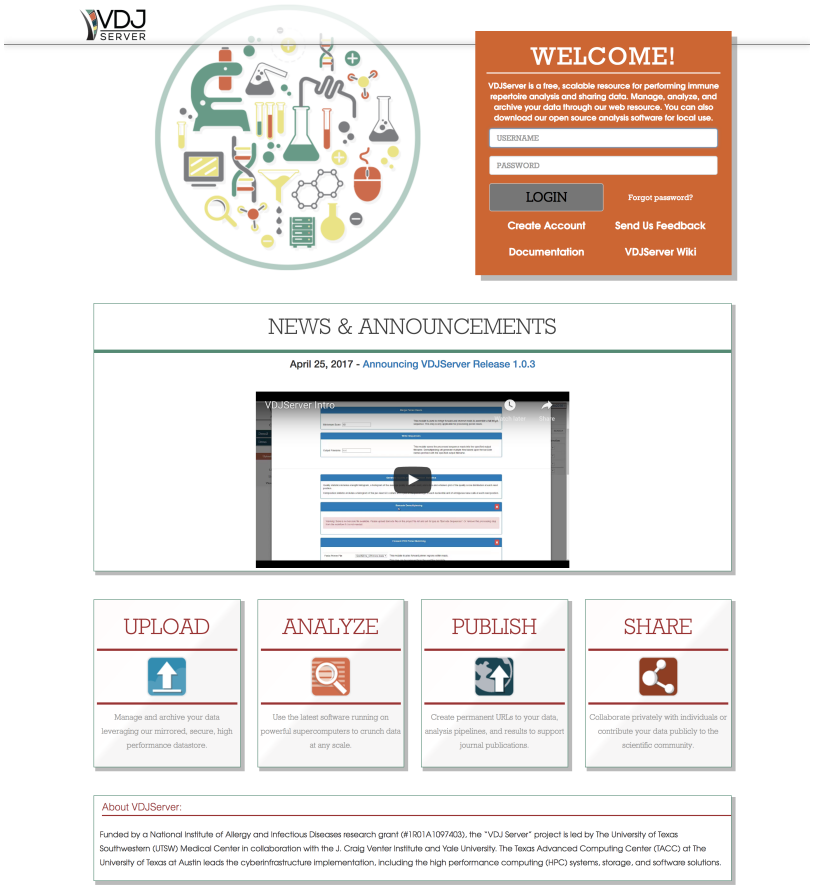
http://www.wikiwand.com/en/Immunological_Genome_Project

https://www.immgen.org

Demo

VDJ Server

- A free, scalable resource for performing immune repertoire analysis and sharing data.
- Manage, analyze, and archive your data through our web resource.
- You can also download our open source analysis software for local use.



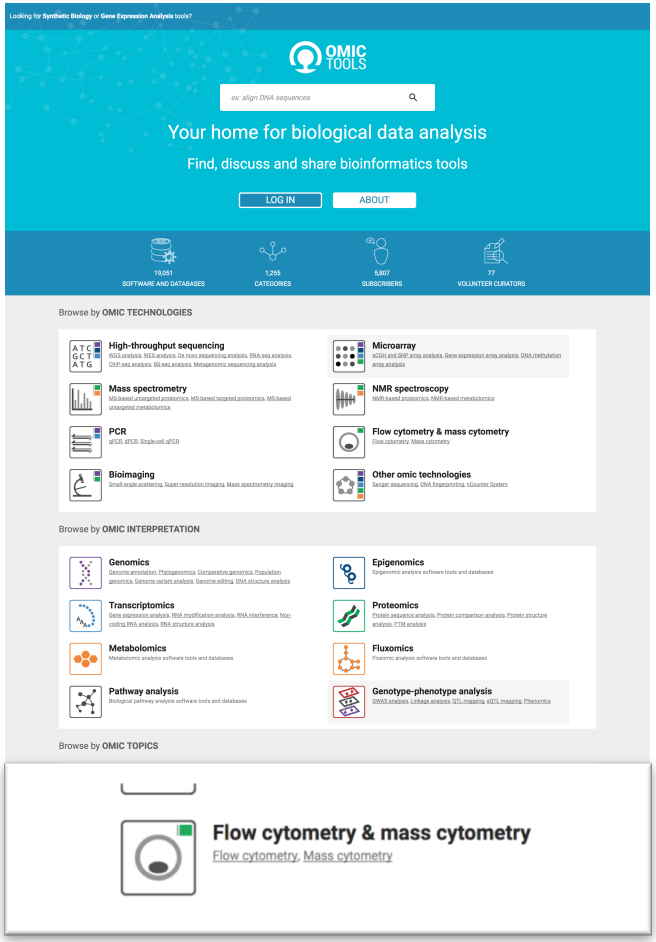
https://vdjserver.org/

Demo

Tool (Software) Repositories

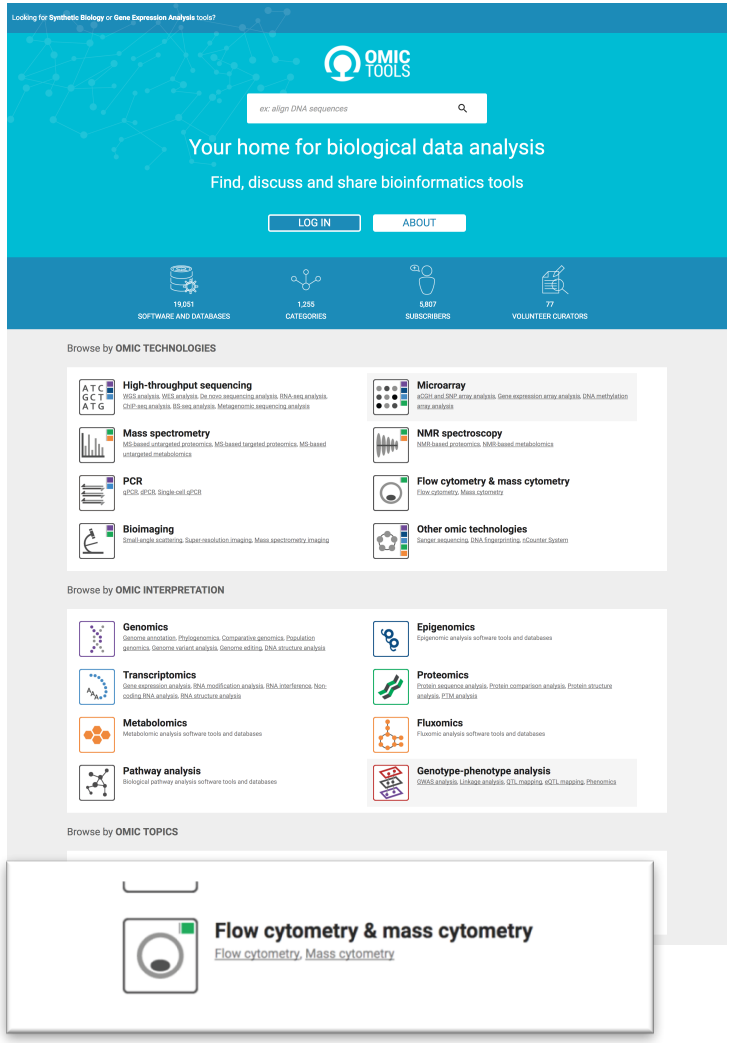
OmicsTools

- Your home for biological data analysis
- Find, discuss and share bioinformatics tools
- 19,051 Software And Databases
- 1,255 Categories
- Browse By
  - Omic Technologies
  - Omic Interpretation
  - Omic Topics



OmicsTools

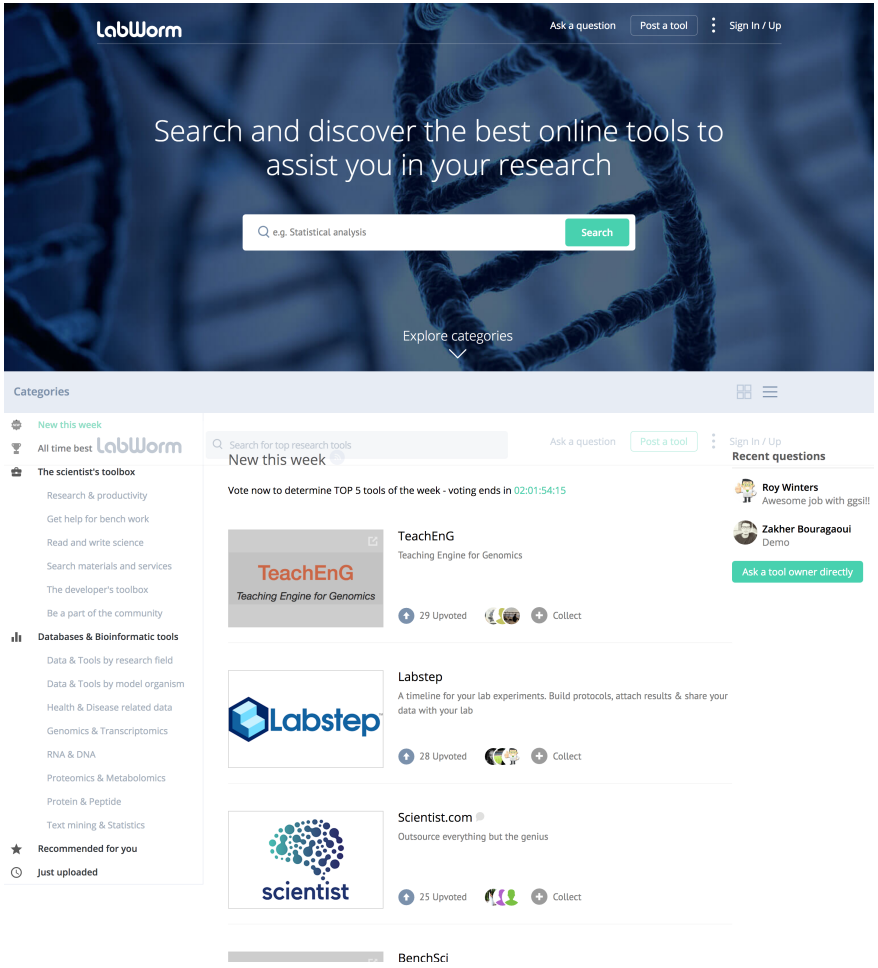
- Omic Technologies
  - Flow cytometry (81 tools)
  - Mass cytometry (15 tools)
- Omic Interpretation
- Omic Topics
  - Immunology



Demo

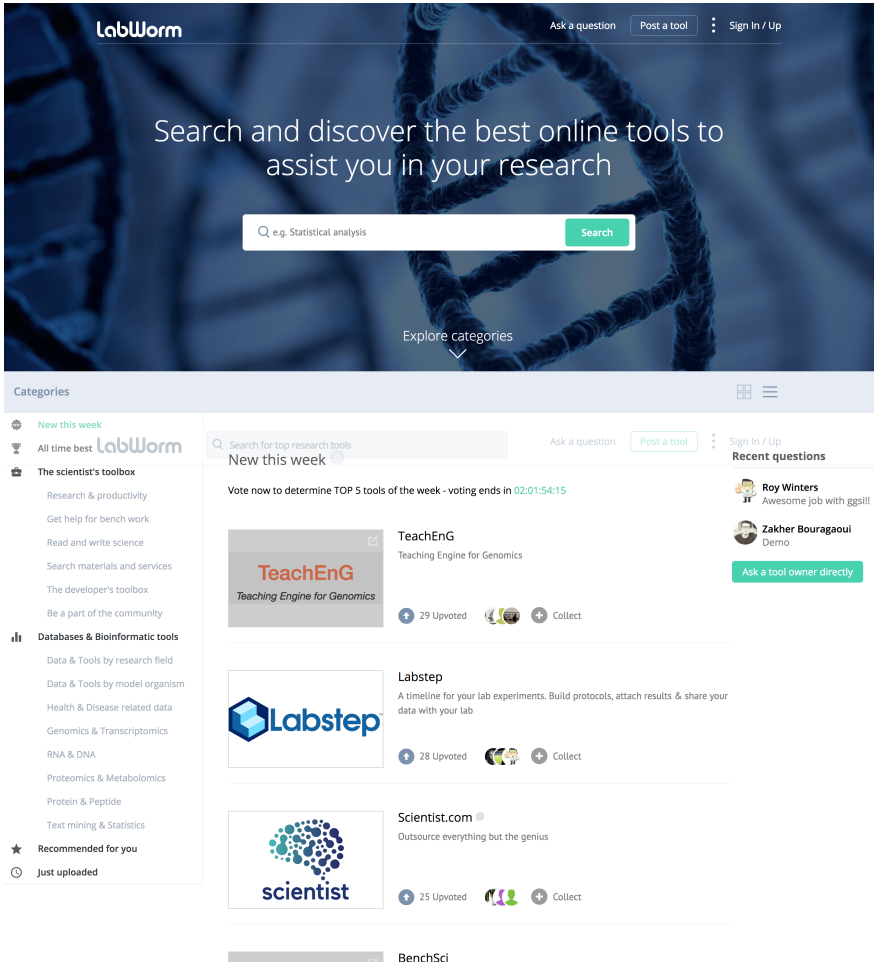
Labworm

- Search and discover the best online tools to assist you in your research
  - The scientist's toolbox
    - Research & productivity
    - Get help for bench work
    - Read and write science
    - Search materials and services
    - The developer's toolbox
    - Be a part of the community
  - Databases & Bioinformatic tools
    - Data & Tools by research field
    - Data & Tools by model organism
    - Health & Disease related data
    - Genomics & Transcriptomics
    - RNA & DNA
    - Proteomics & Metabolomics
    - Protein & Peptide
    - Text mining & Statistics



Labworm - Immunology

- Collections
  - Immunological Tools
    - Margaret Jordan
  - Tools
    - IMGT
    - IMGT/HLA
    - Immgene
    - Immune Epitope Database
    - Kabat Database
    - Immuno Polymorphism Database
    - InnateDB
    - Bcepred
    - INTERFEROME
    - ABCpred
    - Antigenic

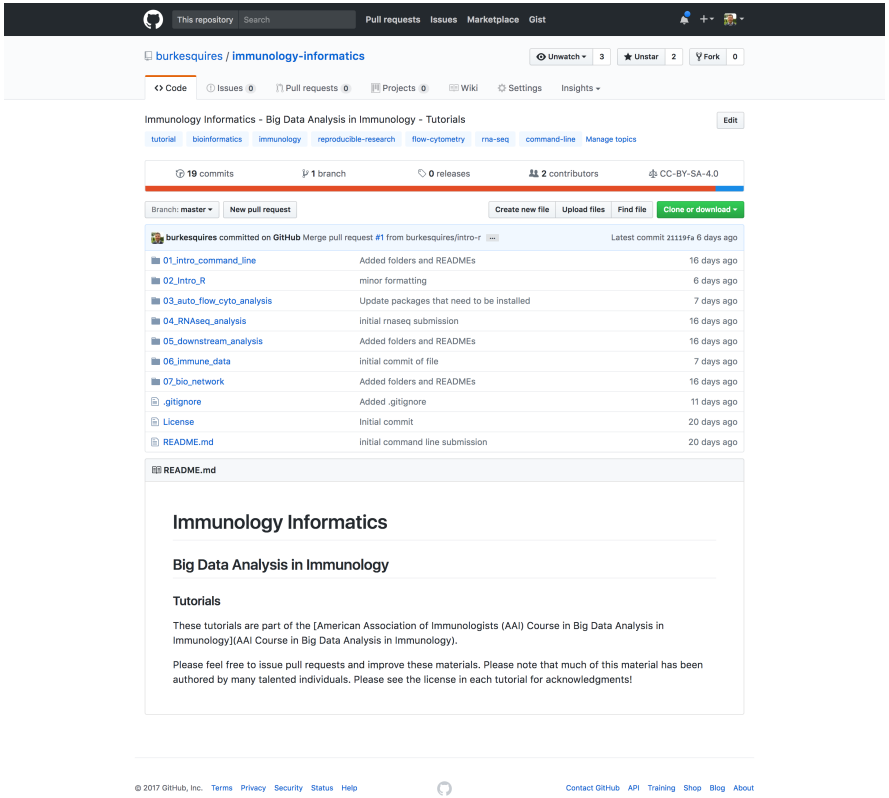


https://labworm.com/collection/Margaret.Jordan/immunological-tools

Demo

GitHub

- GitHub is a development platform inspired by the way you work. From open source to business, you can…
- host and review code
- manage projects
- build software alongside millions of other developers.



Demo

Bioconductor

- Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data.
- Bioconductor uses the R statistical programming language, and is open source and open development.
- It has two releases each year, 1383 software packages, and an active user community.

Demo

Reproducible Research

Becoming a Reproducible Scientist…

- Remember – take it one step at a time
  - Good – Document where you get your data, the steps you use to produce your analysis
  - Better – Script your analysis
  - Best – Scripts your analysis, upload your scripts to GitHub, contribute well document data to a data repository

Additional Resources

Resources

- Big Data 2 Knowledge (BD2K) Lecture series
  - https://bigdatau.ini.usc.edu/data-science-seminars
- Winipedia (via wikiwand)
  - Computational immunology
    - http://www.wikiwand.com/en/Computational_immunology
  - Flow cytometry bioinformatics
    - http://www.wikiwand.com/en/Flow_cytometry_bioinformatics
- “Awesome” Resource Lists of all types (many bioinformatics or computational)
  - https://github.com/sindresorhus/awesome
- Medical clipart – share alike
  - http://smart.servier.com

Resources Used

- BD2K Guide to the Fundamentals of Data Science Series Talks
  - https://www.youtube.com/channel/UCKIDQOa0JcUd3K9C1TS7FLQ
  - Metadata Standards, Dr. Susanna-Assunta Sansone from University of Oxford
  - Introduction to Big Data and the Data Lifecycle, Dr. Mark Musen from Stanford University

Papers of Interest

- Genser, B., Cooper, P. J., Yazdanbakhsh, M., Barreto, M. L. & Rodrigues, L. C. A guide to modern statistical analysis of immunological data. *BMC Immunology 2007 8:1* **8,** 27 (2007).
  - "This paper will help the immunologist to choose the correct statistical approach for a particular research question."
- Zhang, G. L., Sun, J., Chitkushev, L. & Brusic, V. Big Data Analytics in Immunology: A Knowledge-Based Approach. *BioMed Research International* **2014,** 1–9 (2014).

Objectives

- Enable you to name one or more immunologically relevant data standards
- Enable you to search for immunologically relevant data repositories
- Enable you to take a step in becoming a reproducible scientist

Q & A