ViPAR: a software platform for the Virtual Pooling and Analysis of Research data

Associate Professor Kim Carter
Telethon Kids Institute
University of Western Australia

TELETHON KIDS INSTITUTE
Discover. Prevent. Cure.

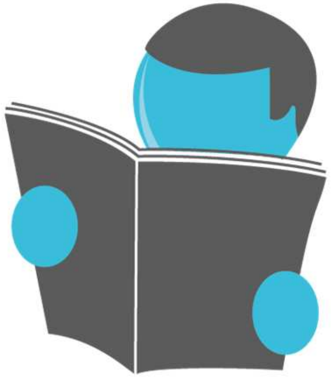
“Any time scientists disagree, it's because we have insufficient data. Then we can agree on what kind of data to get; we get the data; and the data solves the problem.

Either I'm right, or you're right, or we're both wrong. And we move on ... “

Neil deGrasse Tyson

Telethon Kids Institute
Perth, Western Australia

Overview



- Recent media on Data Sharing
- Technologies for Data Sharing
- How did I end up in this space ? (case study)
- Methods for Sharing and Analysing data together
- ViPAR
- Demo

Recent media on data sharing

SHARING CLINICAL RESEARCH DATA

Science as an open enterprise

June 2012

THE ROYAL SOCIETY

OECD Principles and Guidelines for Access to Research Data from Public Funding

OECD

nature

Home | News & Comment | Research | Careers & Jobs | Current Issues | Archive | Audio & V

Archive > Specials & supplements archive > Data Sharing

SPECIALS

See all specials

Open Data White Paper

Unleashing the

#opendata

@uktransparency

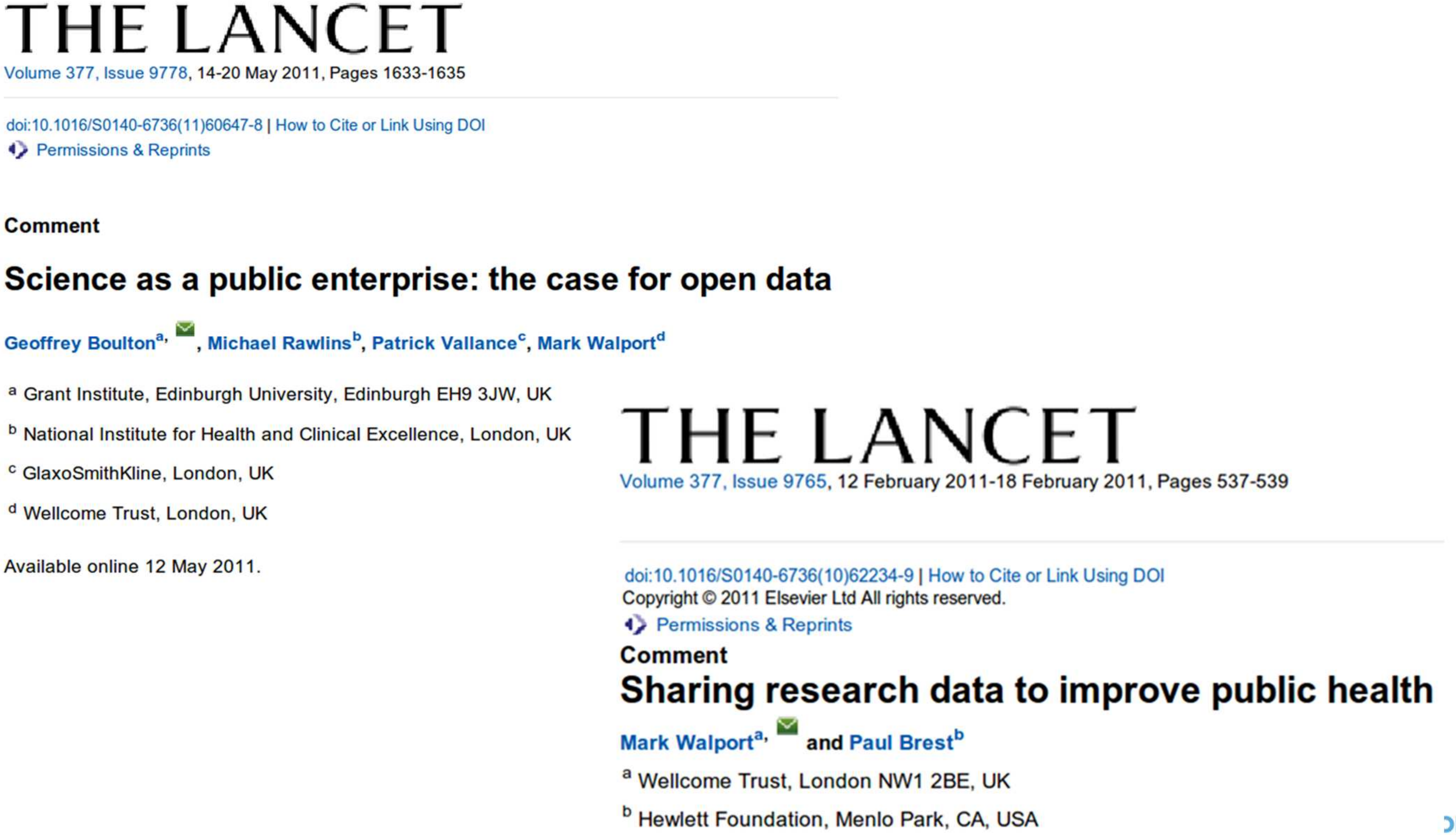
@cabinetofficeuk

"Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property."

RCUK Common Principles on Data Policy

http://www.rcuk.ac.uk/research/datapolicy/

Media on data sharing (cont)



THE LANCET

Volume 377, Issue 9778, 14-20 May 2011, Pages 1633-1635

doi:10.1016/S0140-6736(11)60647-8 | How to Cite or Link Using DOI

Permissions & Reprints

Comment

Science as a public enterprise: the case for open data

Geoffrey Boulton[a], ✉, Michael Rawlins[b], Patrick Vallance[c], Mark Walport[d]

[a] Grant Institute, Edinburgh University, Edinburgh EH9 3JW, UK

[b] National Institute for Health and Clinical Excellence, London, UK

[c] GlaxoSmithKline, London, UK

[d] Wellcome Trust, London, UK

Available online 12 May 2011.

THE LANCET

Volume 377, Issue 9765, 12 February 2011-18 February 2011, Pages 537-539

doi:10.1016/S0140-6736(10)62234-9 | How to Cite or Link Using DOI

Copyright © 2011 Elsevier Ltd All rights reserved.

Permissions & Reprints

Comment

Sharing research data to improve public health

Mark Walport[a], ✉ and Paul Brest[b]

[a] Wellcome Trust, London NW1 2BE, UK

[b] Hewlett Foundation, Menlo Park, CA, USA

Media on data sharing (cont)

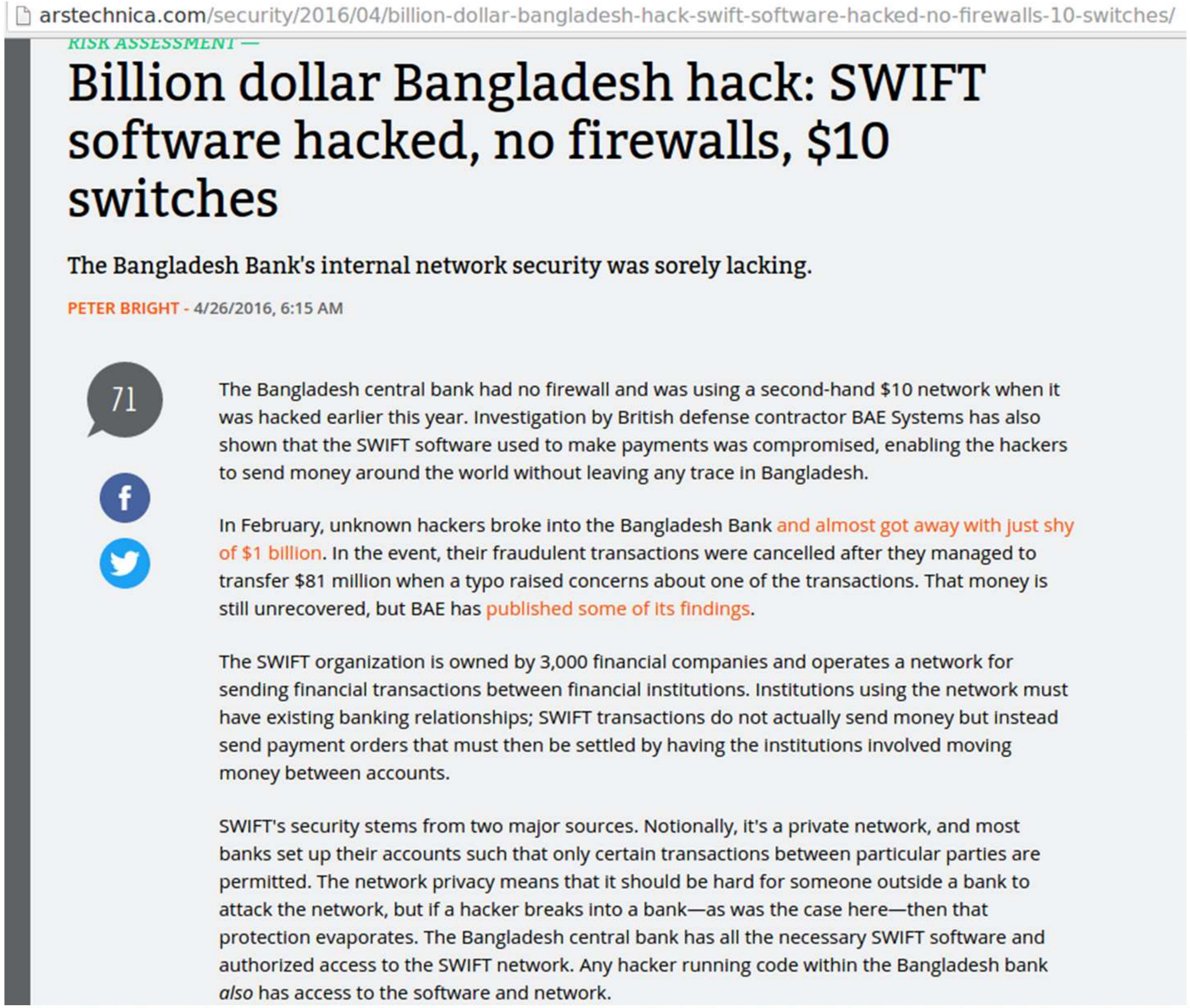
“Increasing availability and promoting efficient data use to maximise public health benefits”

Equitable: balance needs of researchers, communities and funders

Ethical: protect individual privacy and dignity while recognising need to improve health using these data

Efficient: improve quality, value and contribution of research by building on existing, and reducing competition and duplication

Unwanted sharing of data



arstechnica.com/security/2016/04/billion-dollar-bangladesh-hack-swift-software-hacked-no-firewalls-10-switches/

RISK ASSESSMENT —

Billion dollar Bangladesh hack: SWIFT software hacked, no firewalls, $10 switches

The Bangladesh Bank's internal network security was sorely lacking.

PETER BRIGHT - 4/26/2016, 6:15 AM

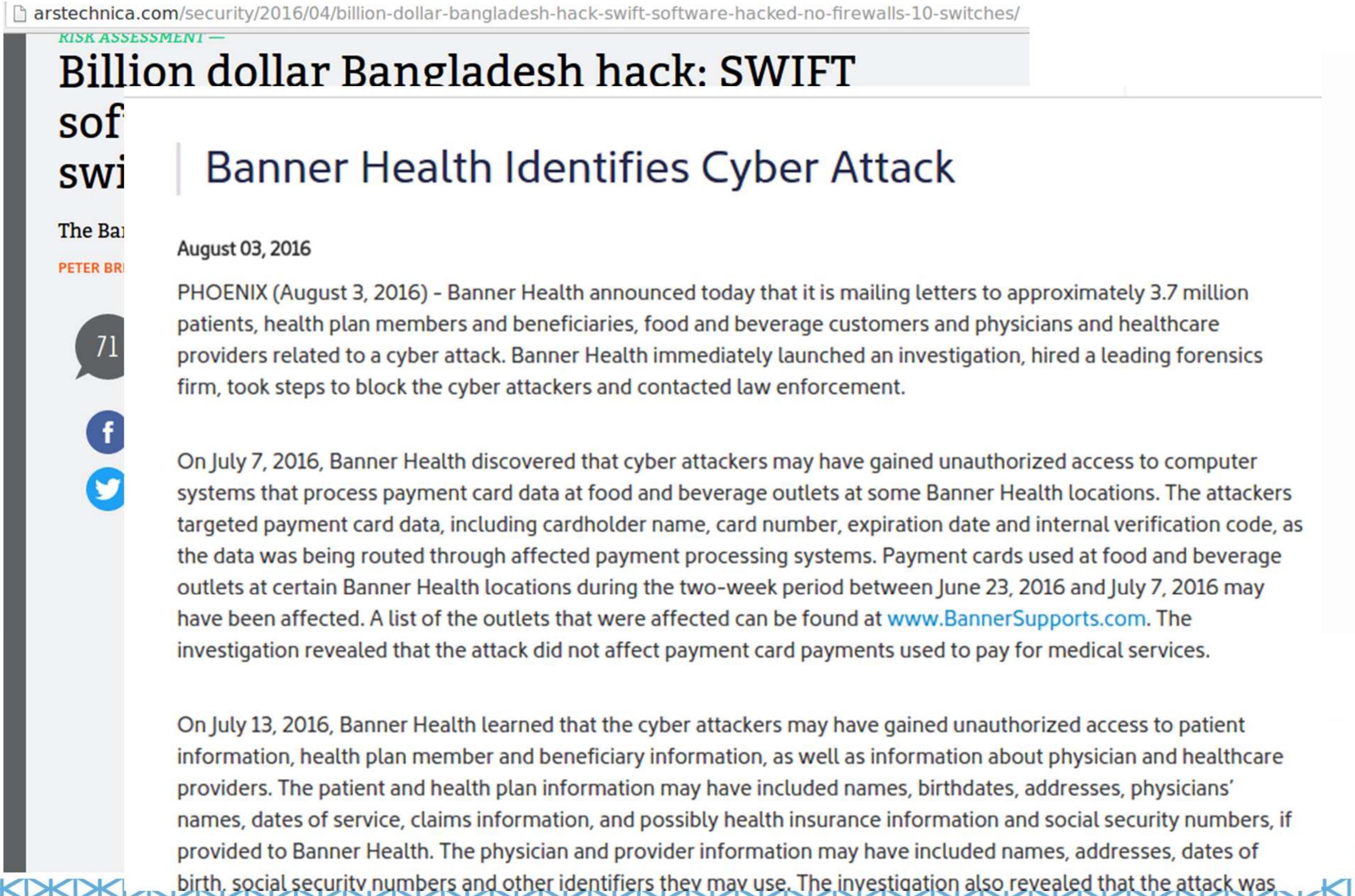
The Bangladesh central bank had no firewall and was using a second-hand $10 network when it was hacked earlier this year. Investigation by British defense contractor BAE Systems has also shown that the SWIFT software used to make payments was compromised, enabling the hackers to send money around the world without leaving any trace in Bangladesh.

In February, unknown hackers broke into the Bangladesh Bank and almost got away with just shy of $1 billion. In the event, their fraudulent transactions were cancelled after they managed to transfer $81 million when a typo raised concerns about one of the transactions. That money is still unrecovered, but BAE has published some of its findings.

The SWIFT organization is owned by 3,000 financial companies and operates a network for sending financial transactions between financial institutions. Institutions using the network must have existing banking relationships; SWIFT transactions do not actually send money but instead send payment orders that must then be settled by having the institutions involved moving money between accounts.

SWIFT's security stems from two major sources. Notionally, it's a private network, and most banks set up their accounts such that only certain transactions between particular parties are permitted. The network privacy means that it should be hard for someone outside a bank to attack the network, but if a hacker breaks into a bank—as was the case here—then that protection evaporates. The Bangladesh central bank has all the necessary SWIFT software and authorized access to the SWIFT network. Any hacker running code within the Bangladesh bank *also* has access to the software and network.

Unwanted sharing of data



arstechnica.com/security/2016/04/billion-dollar-bangladesh-hack-swift-software-hacked-no-firewalls-10-switches/

RISK ASSESSMENT —

Billion dollar Bangladesh hack: SWIFT sof swi

The Ba

PETER BR

71

Banner Health Identifies Cyber Attack

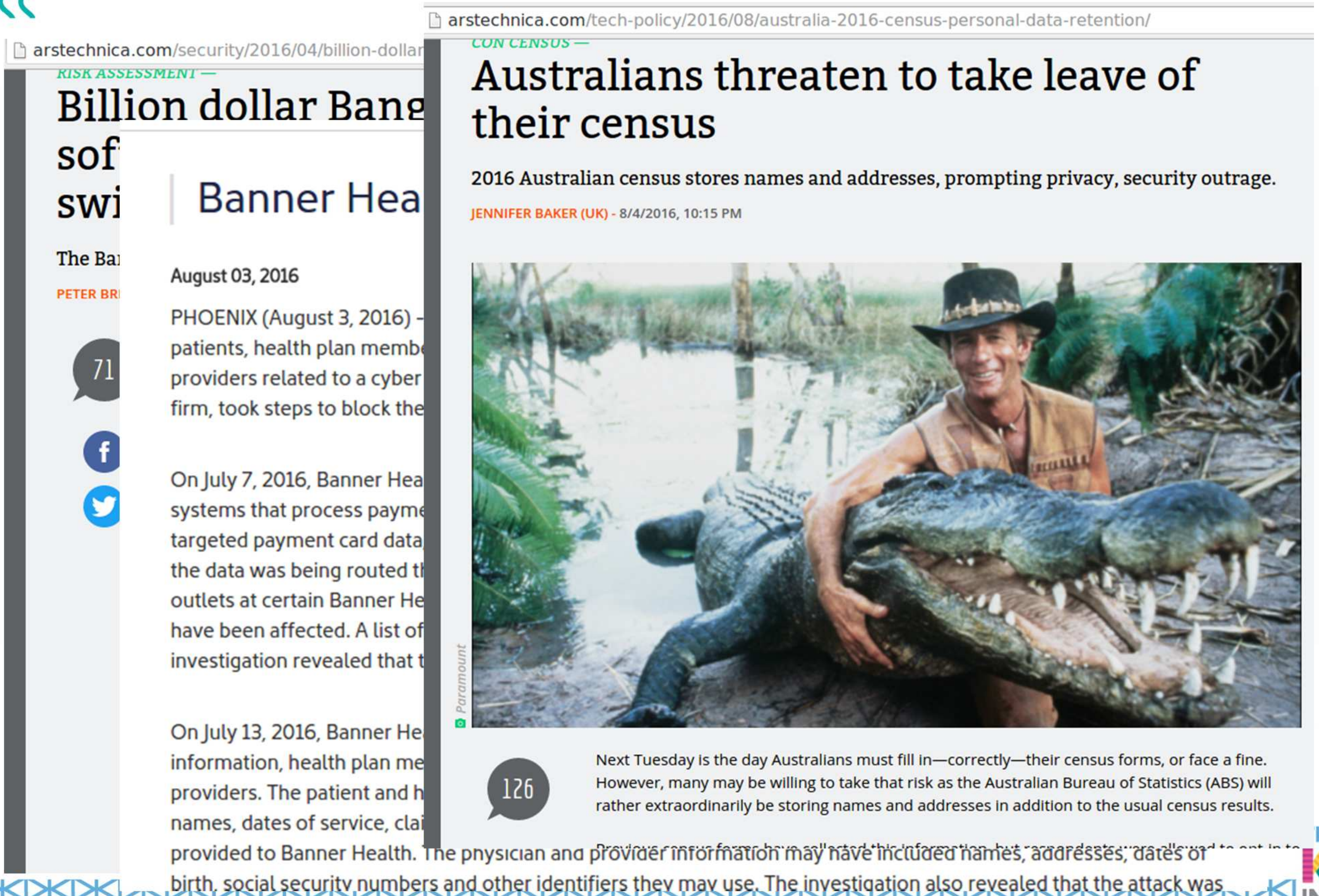
August 03, 2016

PHOENIX (August 3, 2016) - Banner Health announced today that it is mailing letters to approximately 3.7 million patients, health plan members and beneficiaries, food and beverage customers and physicians and healthcare providers related to a cyber attack. Banner Health immediately launched an investigation, hired a leading forensics firm, took steps to block the cyber attackers and contacted law enforcement.

On July 7, 2016, Banner Health discovered that cyber attackers may have gained unauthorized access to computer systems that process payment card data at food and beverage outlets at some Banner Health locations. The attackers targeted payment card data, including cardholder name, card number, expiration date and internal verification code, as the data was being routed through affected payment processing systems. Payment cards used at food and beverage outlets at certain Banner Health locations during the two-week period between June 23, 2016 and July 7, 2016 may have been affected. A list of the outlets that were affected can be found at www.BannerSupports.com. The investigation revealed that the attack did not affect payment card payments used to pay for medical services.

On July 13, 2016, Banner Health learned that the cyber attackers may have gained unauthorized access to patient information, health plan member and beneficiary information, as well as information about physician and healthcare providers. The patient and health plan information may have included names, birthdates, addresses, physicians' names, dates of service, claims information, and possibly health insurance information and social security numbers, if provided to Banner Health. The physician and provider information may have included names, addresses, dates of birth, social security numbers and other identifiers they may use. The investigation also revealed that the attack was

Unwanted sharing of data



arstechnica.com/security/2016/04/billion-dollar

RISK ASSESSMENT —

Billion dollar Bang

sof

swi

The Ba

PETER BR

71

Banner Hea

August 03, 2016

PHOENIX (August 3, 2016) -
patients, health plan membe
providers related to a cyber
firm, took steps to block the

On July 7, 2016, Banner Hea
systems that process payme
targeted payment card data
the data was being routed t
outlets at certain Banner He
have been affected. A list of
investigation revealed that t

On July 13, 2016, Banner He
information, health plan me
providers. The patient and h
names, dates of service, clai
provided to Banner Health. The physician and provider information may have included names, addresses, dates of birth, social security numbers and other identifiers they may use. The investigation also revealed that the attack was

arstechnica.com/tech-policy/2016/08/australia-2016-census-personal-data-retention/

CON CENSUS —

Australians threaten to take leave of their census

2016 Australian census stores names and addresses, prompting privacy, security outrage.

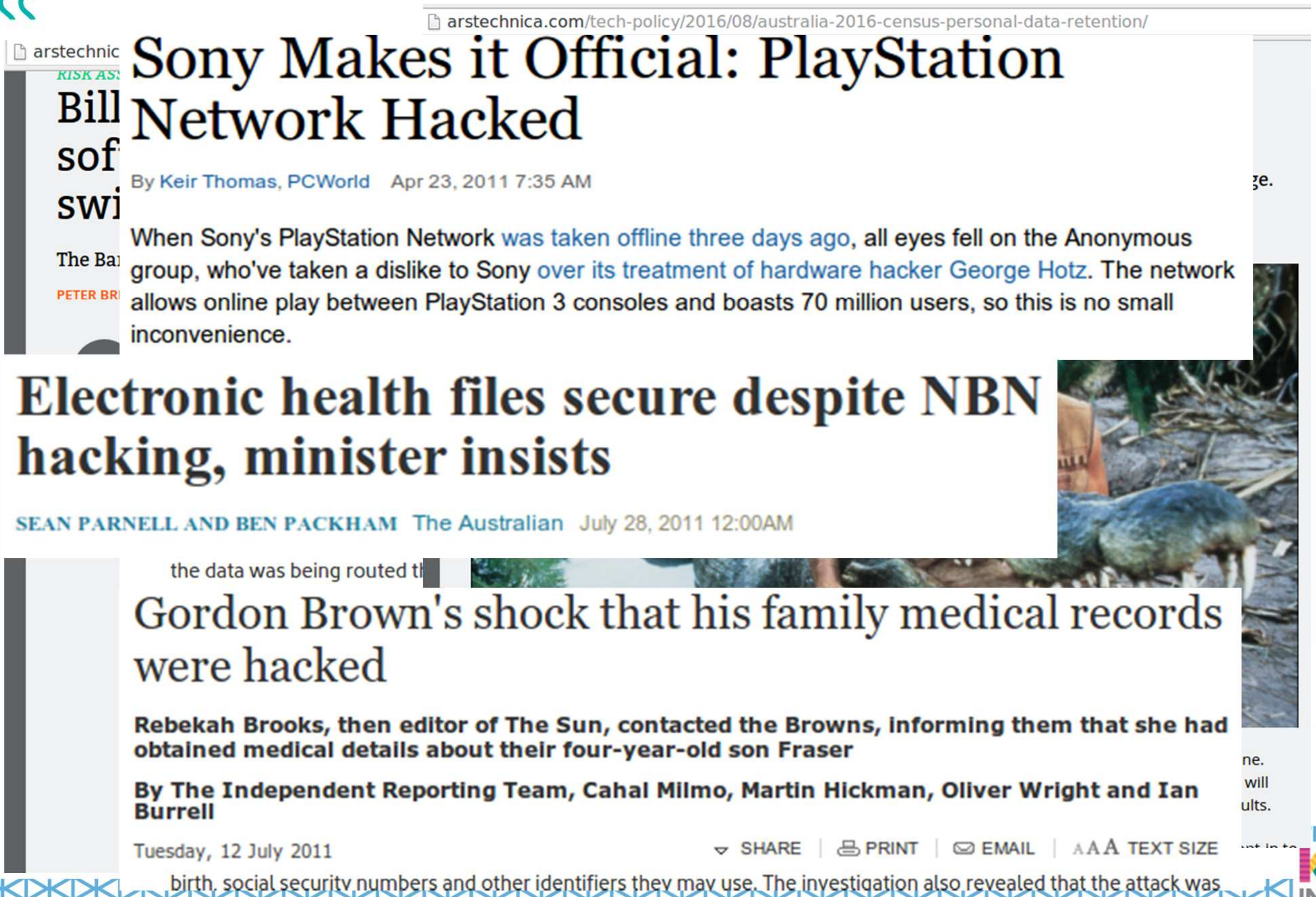
JENNIFER BAKER (UK) - 8/4/2016, 10:15 PM

Paramount

126

Next Tuesday is the day Australians must fill in—correctly—their census forms, or face a fine. However, many may be willing to take that risk as the Australian Bureau of Statistics (ABS) will rather extraordinarily be storing names and addresses in addition to the usual census results.

Unwanted sharing of data



arstechnica.com/tech-policy/2016/08/australia-2016-census-personal-data-retention/

arstechnic

RISK AS

Bill
sof
swi

The Ba

PETER BR

ge.

Sony Makes it Official: PlayStation Network Hacked

By Keir Thomas, PCWorld Apr 23, 2011 7:35 AM

When Sony's PlayStation Network was taken offline three days ago, all eyes fell on the Anonymous group, who've taken a dislike to Sony over its treatment of hardware hacker George Hotz. The network allows online play between PlayStation 3 consoles and boasts 70 million users, so this is no small inconvenience.

Electronic health files secure despite NBN hacking, minister insists

SEAN PARNELL AND BEN PACKHAM The Australian July 28, 2011 12:00AM

the data was being routed t

Gordon Brown's shock that his family medical records were hacked

Rebekah Brooks, then editor of The Sun, contacted the Browns, informing them that she had obtained medical details about their four-year-old son Fraser

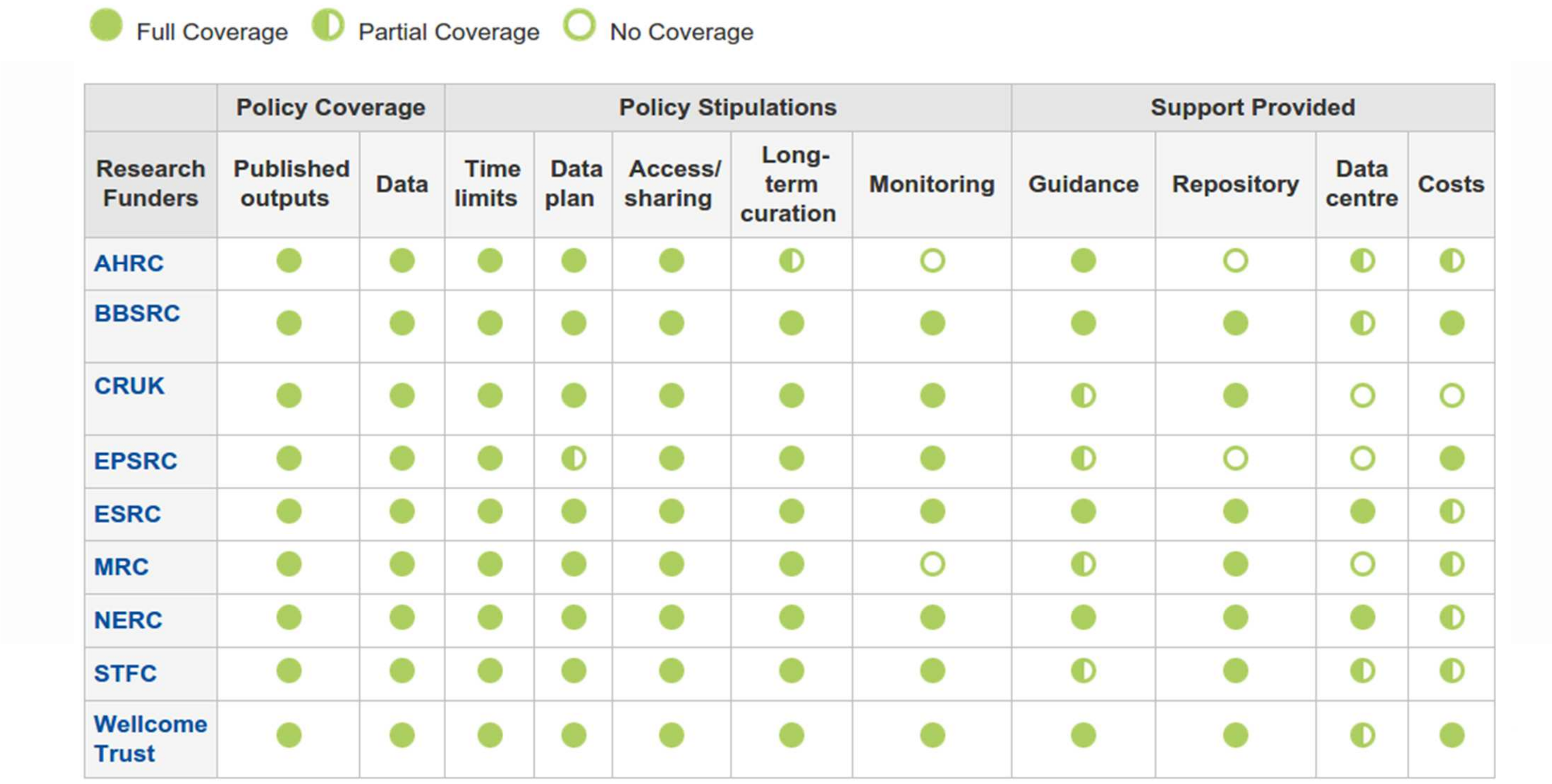
By The Independent Reporting Team, Cahal Milmo, Martin Hickman, Oliver Wright and Ian Burrell

Tuesday, 12 July 2011

ne.
will
ults.

birth, social security numbers and other identifiers they may use. The investigation also revealed that the attack was

Funder expectations (UK)



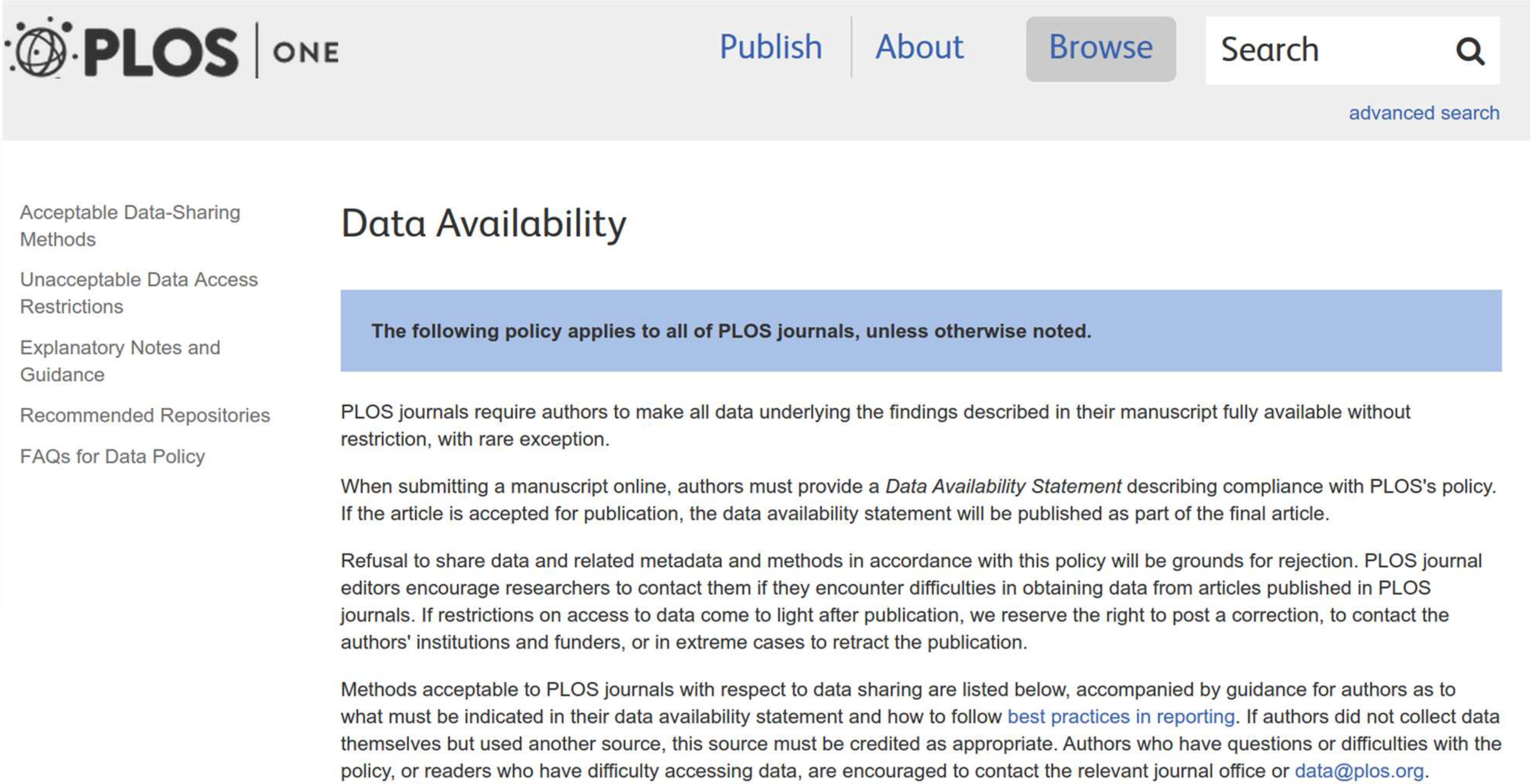
● Full Coverage ◐ Partial Coverage ○ No Coverage

	Policy Coverage		Policy Stipulations					Support Provided			
Research Funders	Published outputs	Data	Time limits	Data plan	Access/ sharing	Long-term curation	Monitoring	Guidance	Repository	Data centre	Costs
AHRC	●	●	●	●	●	◐	○	●	○	◐	◐
BBSRC	●	●	●	●	●	●	●	●	●	◐	●
CRUK	●	●	●	●	●	●	●	◐	●	○	○
EPSRC	●	●	●	◐	●	●	●	◐	○	○	●
ESRC	●	●	●	●	●	●	●	●	●	●	◐
MRC	●	●	●	●	●	●	○	◐	●	○	◐
NERC	●	●	●	●	●	●	●	●	●	●	◐
STFC	●	●	●	●	●	●	●	◐	●	◐	◐
Wellcome Trust	●	●	●	●	●	●	●	●	●	◐	●

http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

Journal expectations

An increasing number of journals expect data to be publicly available at time of publication


PLOS | ONE
Publish
About
Browse
Search
advanced search
Acceptable Data-Sharing Methods
Unacceptable Data Access Restrictions
Explanatory Notes and Guidance
Recommended Repositories
FAQs for Data Policy

Data Availability

The following policy applies to all of PLOS journals, unless otherwise noted.

PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception.

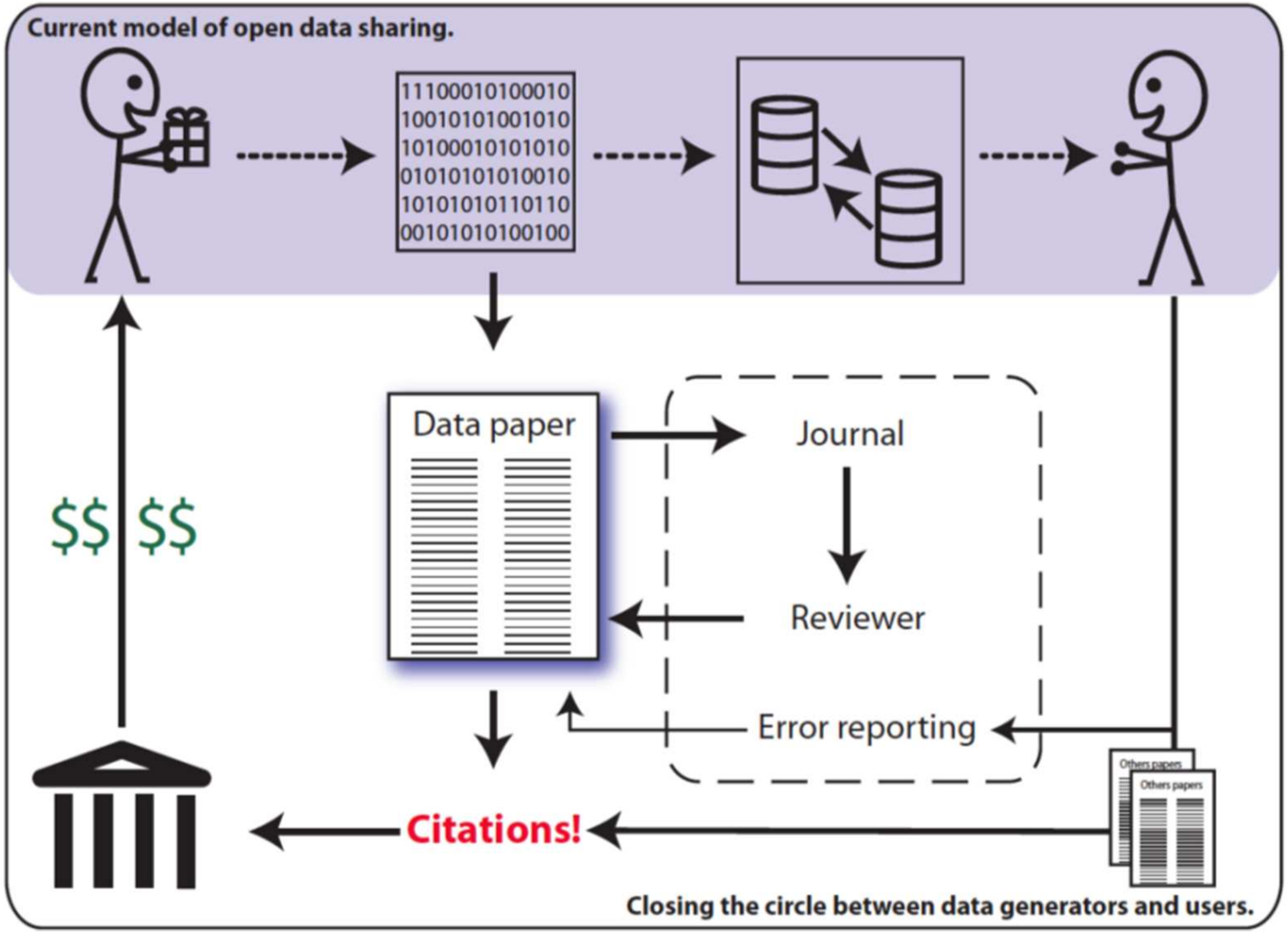
When submitting a manuscript online, authors must provide a *Data Availability Statement* describing compliance with PLOS's policy. If the article is accepted for publication, the data availability statement will be published as part of the final article.

Refusal to share data and related metadata and methods in accordance with this policy will be grounds for rejection. PLOS journal editors encourage researchers to contact them if they encounter difficulties in obtaining data from articles published in PLOS journals. If restrictions on access to data come to light after publication, we reserve the right to post a correction, to contact the authors' institutions and funders, or in extreme cases to retract the publication.

Methods acceptable to PLOS journals with respect to data sharing are listed below, accompanied by guidance for authors as to what must be indicated in their data availability statement and how to follow best practices in reporting. If authors did not collect data themselves but used another source, this source must be credited as appropriate. Authors who have questions or difficulties with the policy, or readers who have difficulty accessing data, are encouraged to contact the relevant journal office or data@plos.org.

http://journals.plos.org/plosone/s/data-availability

Data sharing journals



Gorgolewski, Milham, and Margulies, 2013 Front. Neurosci.

TELETHON KIDS INSTITUTE
Discover. Prevent. Cure.

Large Consortia

UK 10K
RARE GENETIC VARIANTS IN HEALTH AND DISEASE

Wellcome Trust Case Control Consortium

Largest ever study of genetics of common diseases published.

P3G

- To elucidate subtle genetic & environmental effect on common diseases → larger sample sizes req'd

GENOMEUTWIN

- Power comes from the pooled sample size

  (ie an individual level meta-analysis

FANTOM5
FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

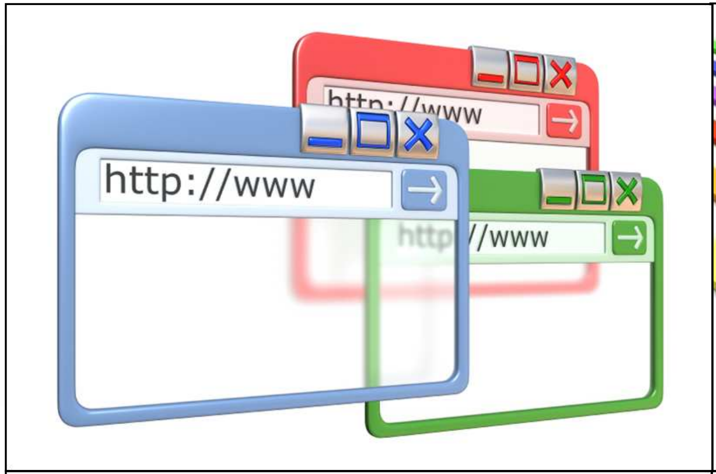
- Multi-site, multi-national means potential ethical, legal and privacy barriers to sharing research data

Disincentives to sharing?

- Policies aren't always compelling
- Labor not always recognised / rewarded
- Authorship: Nominally only First / Last really counts
- Tenure: Metrics Reward Individual not teams
- HIPAA and similar: Scary and significant individual penalties for data loss
- Deidentification: Is it truly possible?

Different Perspectives and Drivers

Technological



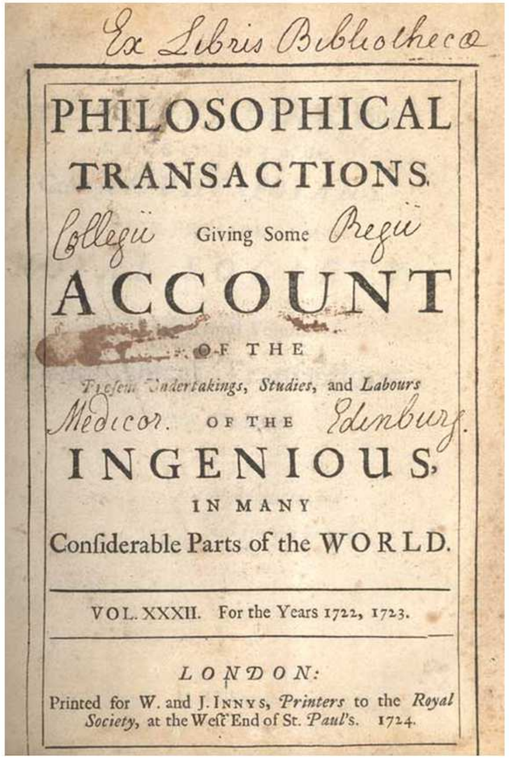
Economic

Legal

Personal

How do you actually share data?

Data sharing by physical media

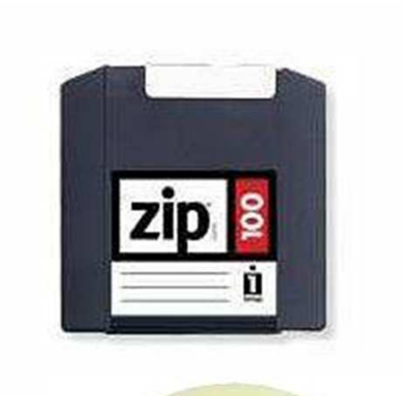


Ex Libris Bibliotheca

PHILOSOPHICAL
TRANSACTIONS
Collegii Giving Some Regii
ACCOUNT
OF THE
Preſent Undertakings, Studies, and Labours
Medicor. OF THE Edinburg.
INGENIOUS,
IN MANY
Conſiderable Parts of the WORLD.

VOL. XXXII. For the Years 1722, 1723.

LONDON:
Printed for W. and J. Innys, Printers to the Royal
Society, at the Weſt End of St. Paul's. 1724.



Usually require local collaborator (proximity) for transfer (manual) or has potentially significant time delays (eg postage)

Technologies for sharing data (by electronic means)

hotfile
EASY SHARE
FileDude
Z-iMu.com
M
ZSHARE
MediaFire
RAPIDSHARE
fileserve
4 shared
FREE FILE SHARING

Removes the barrier of geography to allow shared "storage" of data. Doesn't necessarily make it any easier to do anything once you have the data

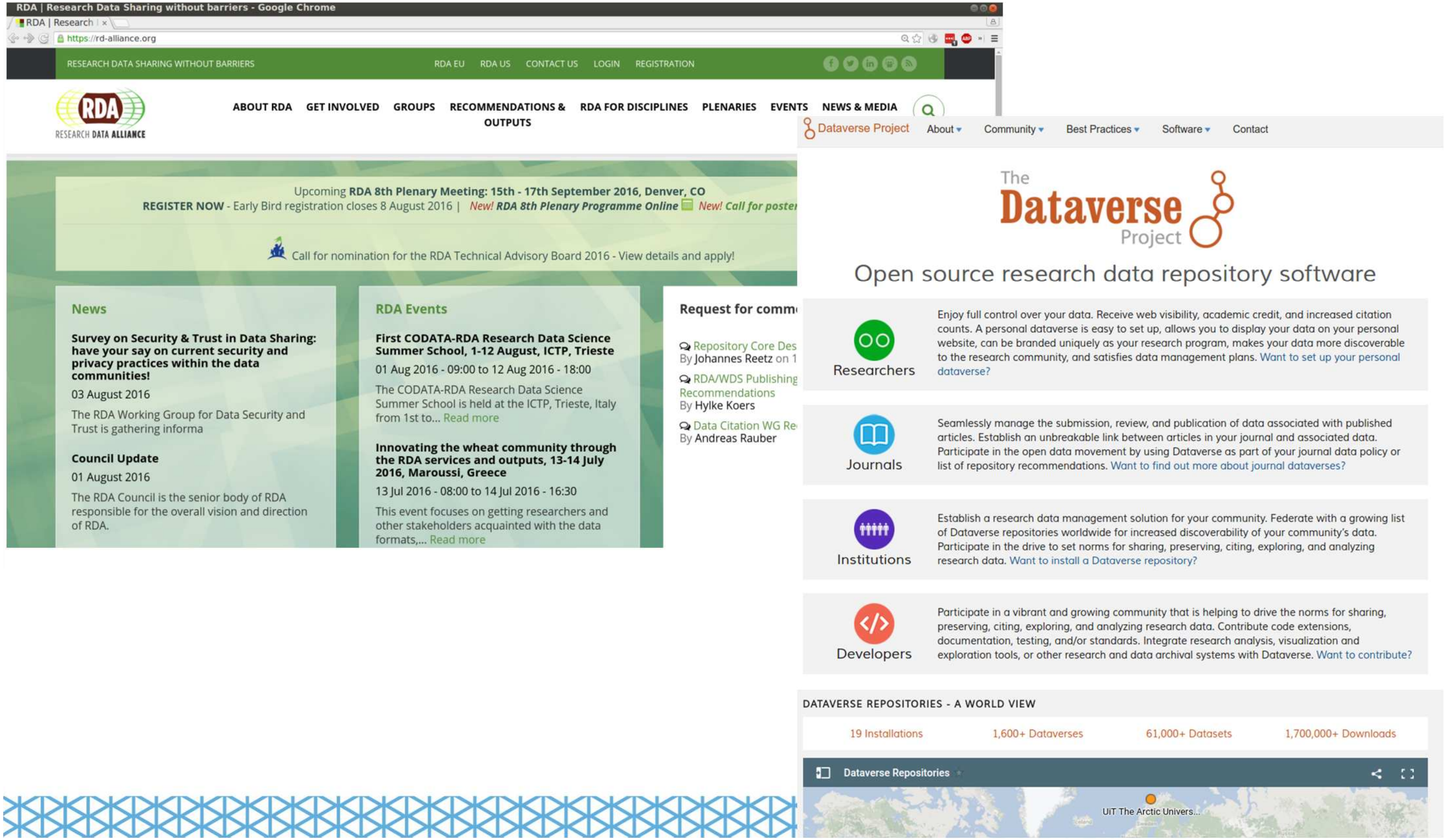
Examples of projects facilitating sharing of research data (Aus)

ARCS – provides data storage, data transfer, collaboration and conferencing facilities

Biogrid – platform for integrating and analysing clinical, imaging and biospecimen data across jurisdictions

NeCTAR – cloud services for workflows, tools and servers

Examples of successful data sharing meta-resources



Upcoming **RDA 8th Plenary Meeting: 15th - 17th September 2016, Denver, CO**

REGISTER NOW - Early Bird registration closes 8 August 2016 | *New!* ***RDA 8th Plenary Programme Online*** *New! Call for poster*

Call for nomination for the RDA Technical Advisory Board 2016 - View details and apply!

News

Survey on Security & Trust in Data Sharing: have your say on current security and privacy practices within the data communities!

03 August 2016

The RDA Working Group for Data Security and Trust is gathering informa

Council Update

01 August 2016

The RDA Council is the senior body of RDA responsible for the overall vision and direction of RDA.

RDA Events

First CODATA-RDA Research Data Science Summer School, 1-12 August, ICTP, Trieste

01 Aug 2016 - 09:00 to 12 Aug 2016 - 18:00

The CODATA-RDA Research Data Science Summer School is held at the ICTP, Trieste, Italy from 1st to... Read more

Innovating the wheat community through the RDA services and outputs, 13-14 July 2016, Maroussi, Greece

13 Jul 2016 - 08:00 to 14 Jul 2016 - 16:30

This event focuses on getting researchers and other stakeholders acquainted with the data formats,... Read more

Request for comm

Repository Core Des
By Johannes Reetz on 1

RDA/WDS Publishing Recommendations
By Hylke Koers

Data Citation WG Re
By Andreas Rauber

Open source research data repository software

Researchers

Enjoy full control over your data. Receive web visibility, academic credit, and increased citation counts. A personal dataverse is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable to the research community, and satisfies data management plans. Want to set up your personal dataverse?

Journals

Seamlessly manage the submission, review, and publication of data associated with published articles. Establish an unbreakable link between articles in your journal and associated data. Participate in the open data movement by using Dataverse as part of your journal data policy or list of repository recommendations. Want to find out more about journal dataverses?

Institutions

Establish a research data management solution for your community. Federate with a growing list of Dataverse repositories worldwide for increased discoverability of your community's data. Participate in the drive to set norms for sharing, preserving, citing, exploring, and analyzing research data. Want to install a Dataverse repository?

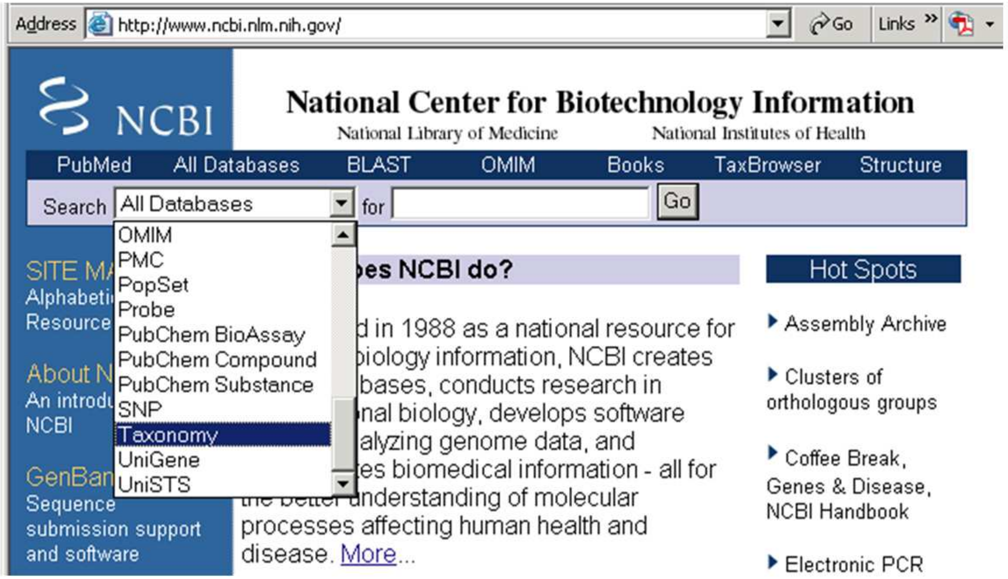
Developers

Participate in a vibrant and growing community that is helping to drive the norms for sharing, preserving, citing, exploring, and analyzing research data. Contribute code extensions, documentation, testing, and/or standards. Integrate research analysis, visualization and exploration tools, or other research and data archival systems with Dataverse. Want to contribute?

DATAVERSE REPOSITORIES - A WORLD VIEW

19 Installations	1,600+ Dataverses	61,000+ Datasets	1,700,000+ Downloads

Examples of successful data sharing projects (repositories)



Address http://www.ncbi.nlm.nih.gov/ Go Links

NCBI

National Center for Biotechnology Information

National Library of Medicine   National Institutes of Health

PubMed | All Databases | BLAST | OMIM | Books | TaxBrowser | Structure

Search All Databases for Go

OMIM
PMC
PopSet
Probe
PubChem BioAssay
PubChem Compound
PubChem Substance
SNP
Taxonomy
UniGene
UniSTS

SITE MA
Alphabeti
Resource

About N
An introdu
NCBI

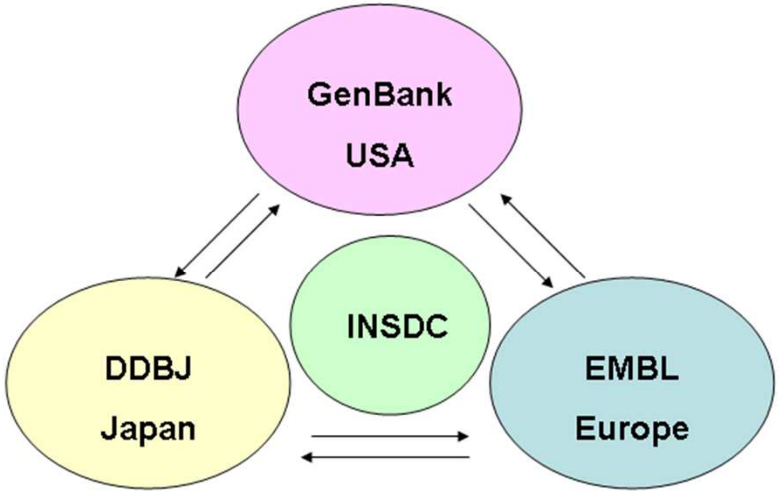
GenBan
Sequence submission support and software

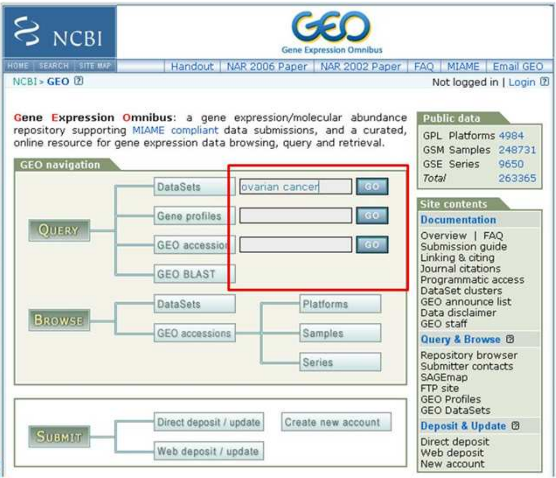
es NCBI do?

d in 1988 as a national resource for biology information, NCBI creates bases, conducts research in nal biology, develops software alyzing genome data, and es biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

Hot Spots

- Assembly Archive
- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR





NCBI

GEO Gene Expression Omnibus

HOME | SEARCH | SITE MAP | Handout | NAR 2006 Paper | NAR 2002 Paper | FAQ | MIAME | Email GEO

NCBI > GEO   Not logged in | Login

Gene Expression Omnibus: a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.

GEO navigation

QUERY: DataSets [ovarian cancer] GO; Gene profiles GO; GEO accession GO; GEO BLAST

BROWSE: DataSets; GEO accessions: Platforms, Samples, Series

SUBMIT: Direct deposit / update; Web deposit / update; Create new account

Public data

GPL	Platforms	4984
GSM	Samples	248731
GSE	Series	9650
Total		263365

Site contents

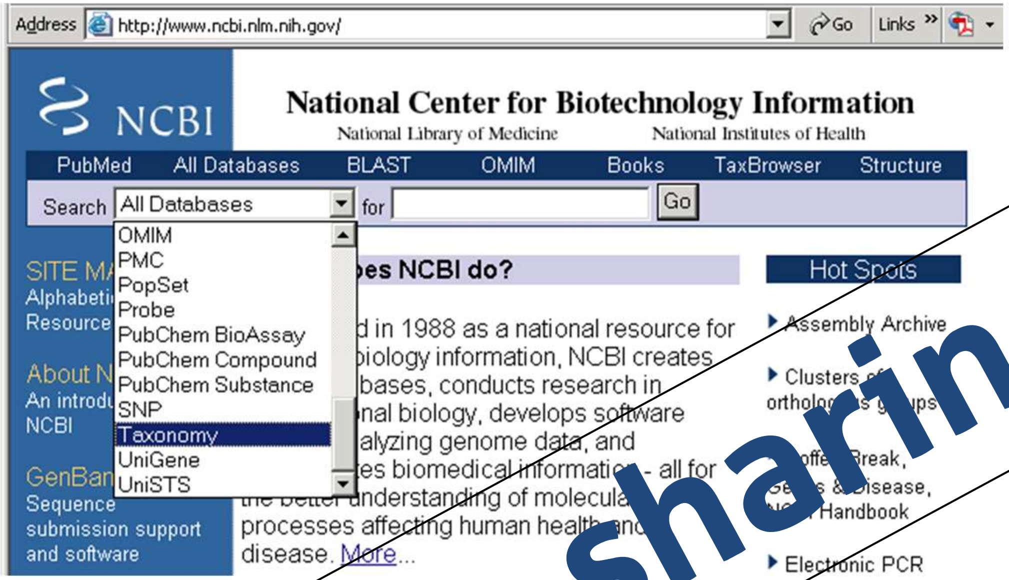
Documentation
Overview | FAQ
Submission guide
Linking & citing
Journal citations
Programmatic access
DataSet clusters
GEO announce list
Data disclaimer
GEO staff

Query & Browse
Repository browser
Submitter contacts
SAGEmap
FTP site
GEO Profiles
GEO DataSets

Deposit & Update
Direct deposit
Web deposit
New account

Examples of successful data sharing projects (repositories)

Data sharing 1.0



National Center for Biotechnology Information

National Library of Medicine National Institutes of Health

Search: All Databases

SITE MA… Alphabeti… Resource…

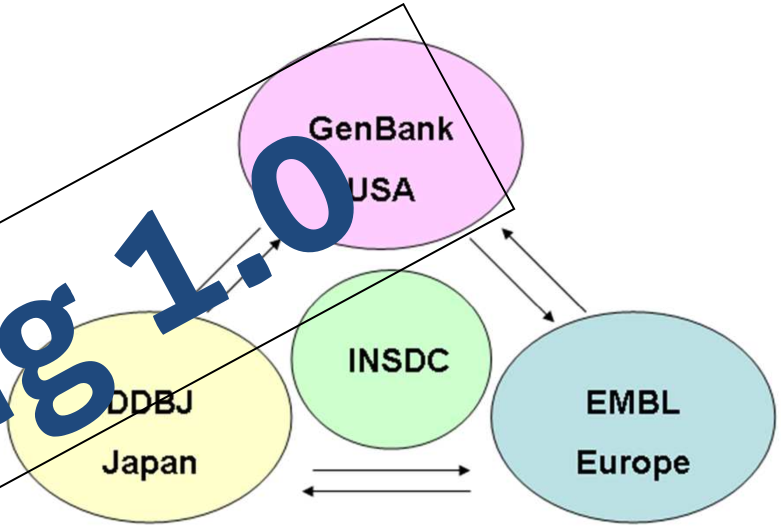
About N… An introdu… NCBI

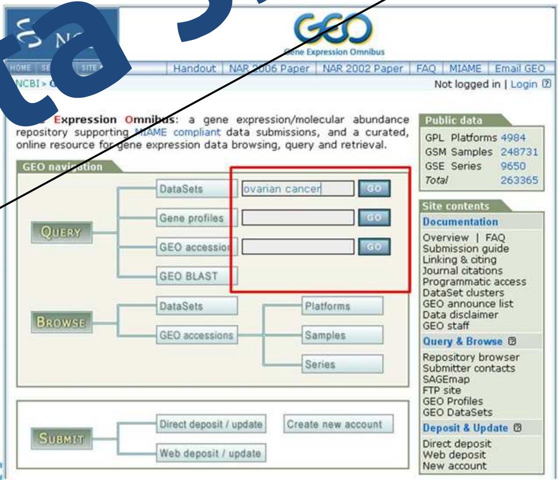
GenBan… Sequence submission support and software

…es NCBI do?

…d in 1988 as a national resource for …biology information, NCBI creates …bases, conducts research in …nal biology, develops software …alyzing genome data, and …es biomedical informatio… - all for the better understanding of molecula… processes affecting human health and disease. More...

…offe… Break, …Ge…s & Disease, NC… Handbook





…e Expression Omnibus: a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.

DataSets: ovarian cancer

Public data

GPL	Platforms	4984
GSM	Samples	248731
GSE	Series	9650
Total		263365

Site contents

Documentation

Overview | FAQ
Submission guide
Linking & citing
Journal citations
Programmatic access
DataSet clusters
GEO announce list
Data disclaimer
GEO staff

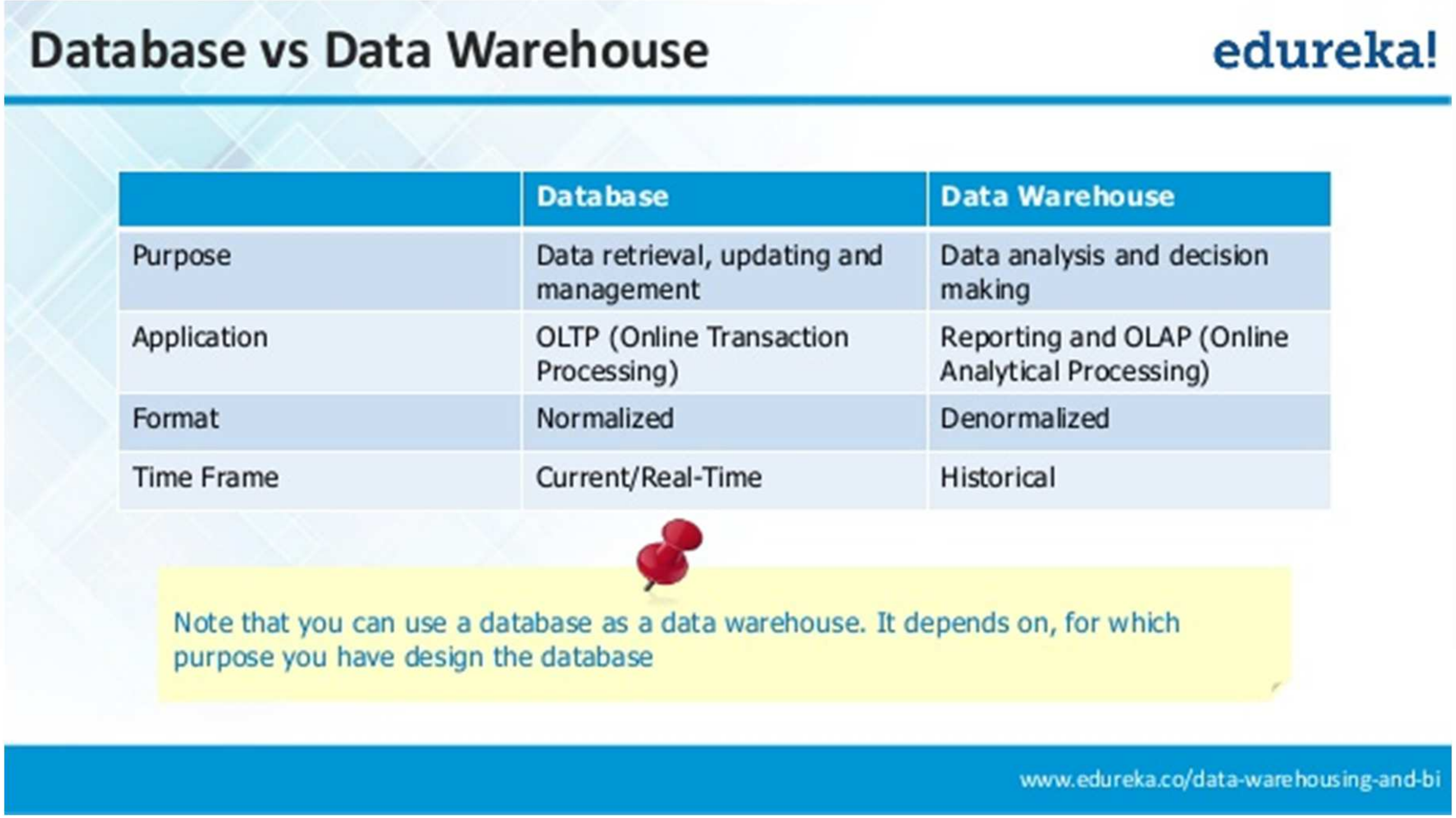
Query & Browse

Repository browser
Submitter contacts
SAGEmap
FTP site
GEO Profiles
GEO DataSets

Deposit & Update

Direct deposit
Web deposit
New account

Data warehouses vs Databases



Database vs Data Warehouse

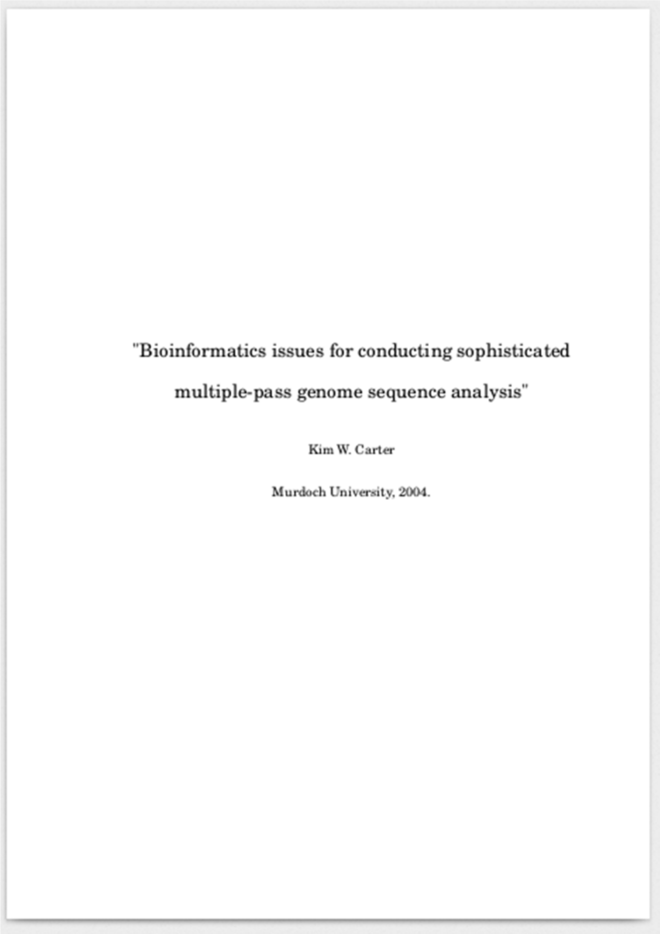
edureka!

	Database	Data Warehouse
Purpose	Data retrieval, updating and management	Data analysis and decision making
Application	OLTP (Online Transaction Processing)	Reporting and OLAP (Online Analytical Processing)
Format	Normalized	Denormalized
Time Frame	Current/Real-Time	Historical

Note that you can use a database as a data warehouse. It depends on, for which purpose you have design the database

www.edureka.co/data-warehousing-and-bi

Data warehouses vs Databases



"Bioinformatics issues for conducting sophisticated multiple-pass genome sequence analysis"

Kim W. Carter

Murdoch University, 2004.

How/why did I get interested in data sharing technologies?

Fig. 5. The Side Step by which an Opponent is Eluded.

The International Consortium for Autism Registry Epidemiology (iCARE)

iCARE annual meeting, Sweden 2011

Vision for iCARE and autism research

Motivated and willing collaborators wanted to join forces to create a resource that allows for analyses that:

- cannot be performed with a single existing system
- enhance the analytic potential of a single data system
- allow direct comparison of findings across different data systems with eg geographic, population, or data collection variation.

Overarching aims of iCARE

1. **Funded (Autism Speaks 2009-2013) to setup a <u>multi-national virtual data set</u> for the study of pre- and perinatal risk factors for autism.**
    - Establish a minimal database requirement.
    - Establish and conduct data preparation protocols across registries.
    - Enable and test centralized data access protocols to multiple registries.
    - Establish a collaborative framework and guidelines for a working relationship among sites/investigators.

2. **Demonstrate the utility of the resource**

How did I get involved with iCARE?

Setup a **multi-national virtual data set** for the study of pre- and perinatal risk factors for autism.

The group had funding and ethics approval for virtual pooling but the original IT group had to withdraw

iCARE consortium characteristics

Site	Population Size	Birth Years	Births/Year	Coverage	Health Care Provision
Denmark	5.5 million	1980-2007	62,000	National	Public
Finland	5.4 million	1987-2008	60,000	National	Public
Israel	7.6 million	1987-2006	86,000	National	Public
Norway	4.8 million	1980-2005	55,000	National	Public
Sweden	9.4 million	1980-2008	107,000	National	Public
Western Australia	1.9 million	1983-1999	24,000	State	Public and private

Population-based registries → Large samples, unbiased, prospective, with ability to link with other population databases

Data dictionary and harmonisation

Critical to the success of the project (and any data sharing project) is making data comparable across sites

Challenge

Diverse data availability and formats, over time and by site

Solution: Harmonisation

generic term for procedures that create comparability between data derived from different sources

Data dictionary and harmonisation

Minimum Variable List

Data Availability Surveys

Data Dictionary

Certification

Archives

Data dictionary V1 (Sep 2010) – 48 variables

Data dictionary V2 (Mar 2011) – 52 variables

Data dictionary V3 (Mar 2013) – 58 variables

Recap: So why do we want to share research data?

Benefits

- Increased power, leading to subgroup analyses and interactions
- Ability to compare outcomes and validate models across sites

Pitfalls

- Cost of harmonisation – time and $
- Difficult to validate/error check when anonymous
- Potentially more complex analysis
- Ethico-legal issues with privacy & consent
- Requires strong collaborations

Are there electronic methods that facilitate data sharing and analysis together?

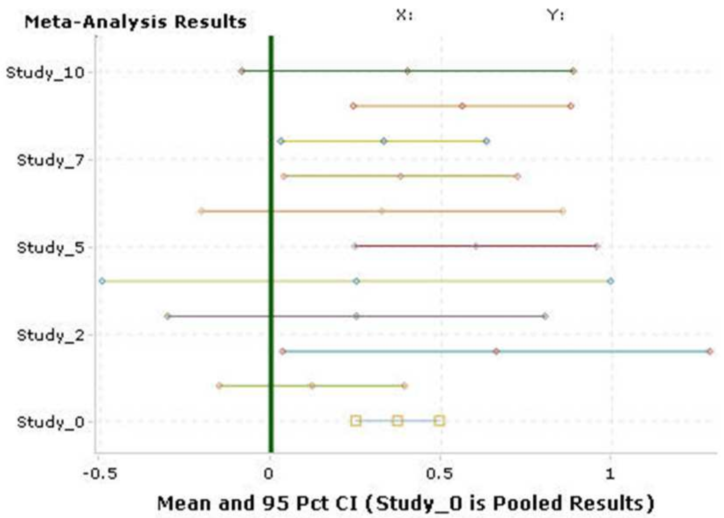
Methods for Sharing and Analysing Research Data

1. Traditional Meta Analysis

Combining existing results of analyses on similar outcomes and predictors using similar methodology

- needs published data on similar outcomes and predictors using similar methodology
- can suffer from “resolution”

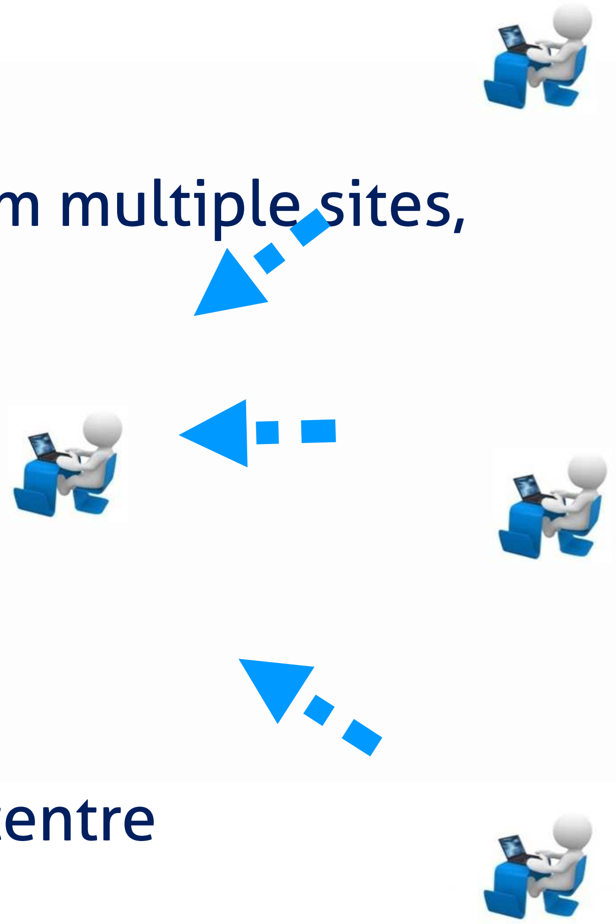
As individual level data may not
 be used



Methods for Sharing and Analysing Research Data (cont)

2. Manual Data Pooling

Analysis of harmonised pooled data from multiple sites, sent to a single analysis site



- (manually) unified methodology
- needs harmonised data
- consent and ethics for sharing
- significant effort for a single analysis centre
- requires strong collaboration

Methods for Sharing and Analysing Research Data (cont)

3. Automated Data Pooling (federation)

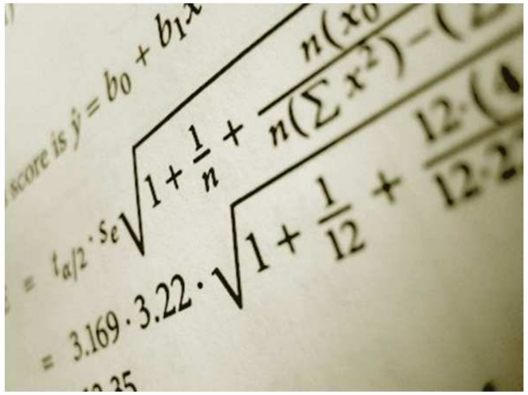
Analysis of harmonised pooled data automatically from multiple sites (pooling in the "cloud")

- unified methodology
- needs harmonised data
- consent and ethics for sharing
- strong collaboration
- requires informatics infrastructure to enable the federation process
- Requires some infrastructure at all data-contributing sites

Methods for Sharing and Analysing Research Data (cont)

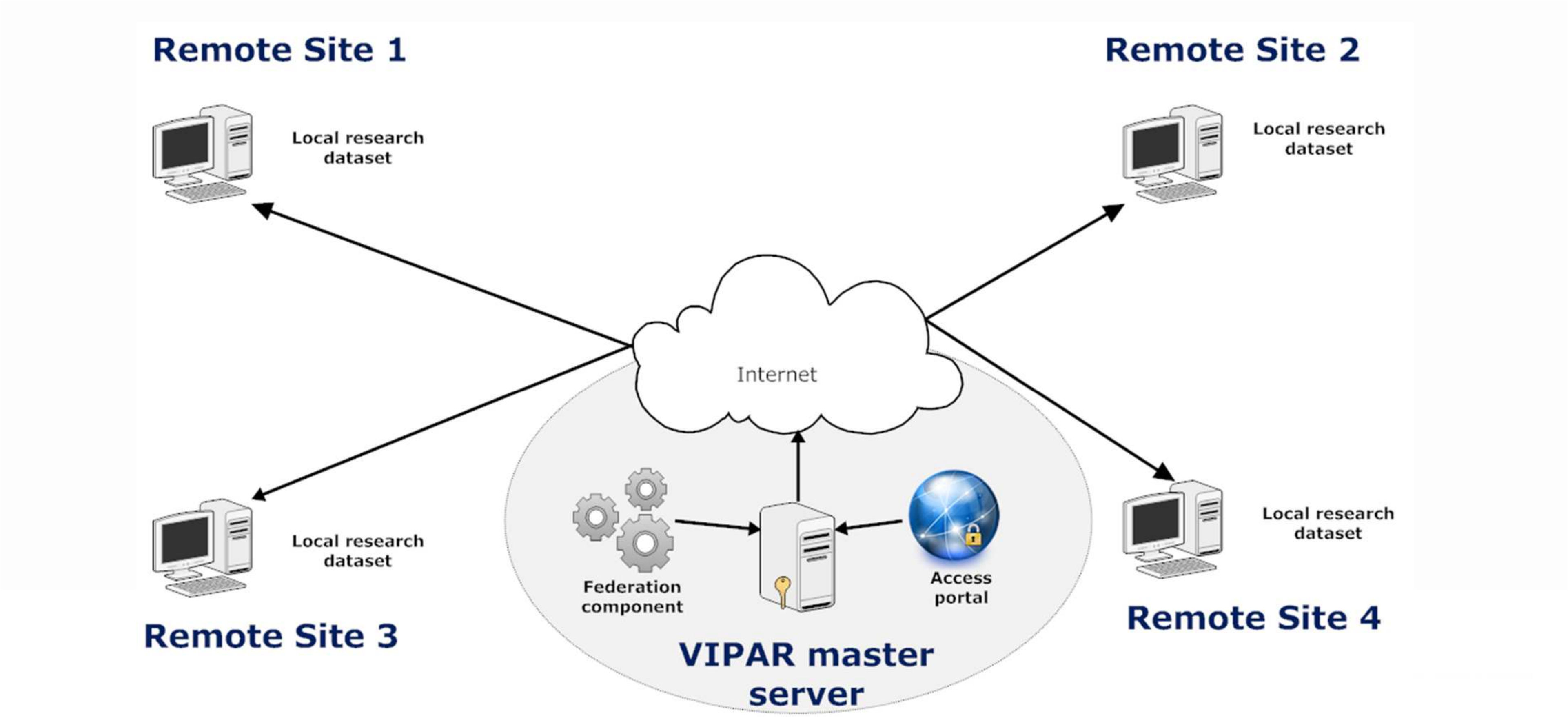
4. Automated decentralised analysis – eg DataSHIELD

Analysis of harmonised data automatically from multiple sites by pooling statistics
(ie no data transfer)

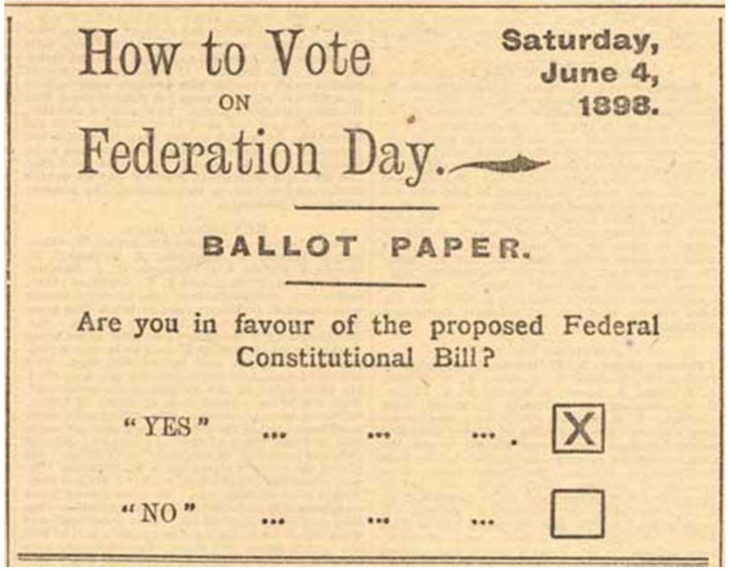


- unified methodology
- needs harmonised data
- may need consent and ethics for sharing
- potentially limited types of analysis
- Requires significant informatics infrastructure at all data-contributing sites

ViPAR: Virtual infrastructure for pooling and analysis of research data



What is Federation?



How to Vote ON Federation Day.

Saturday, June 4, 1898.

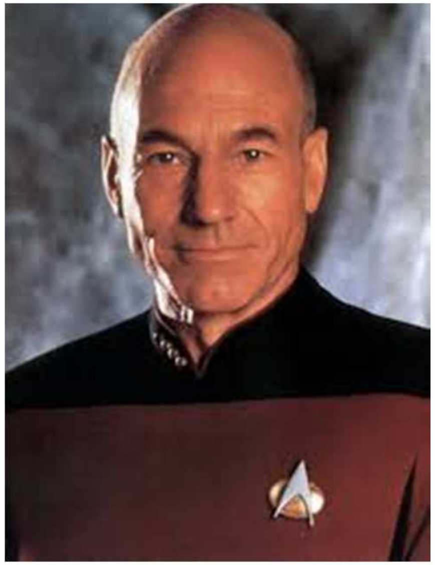
BALLOT PAPER.

Are you in favour of the proposed Federal Constitutional Bill?

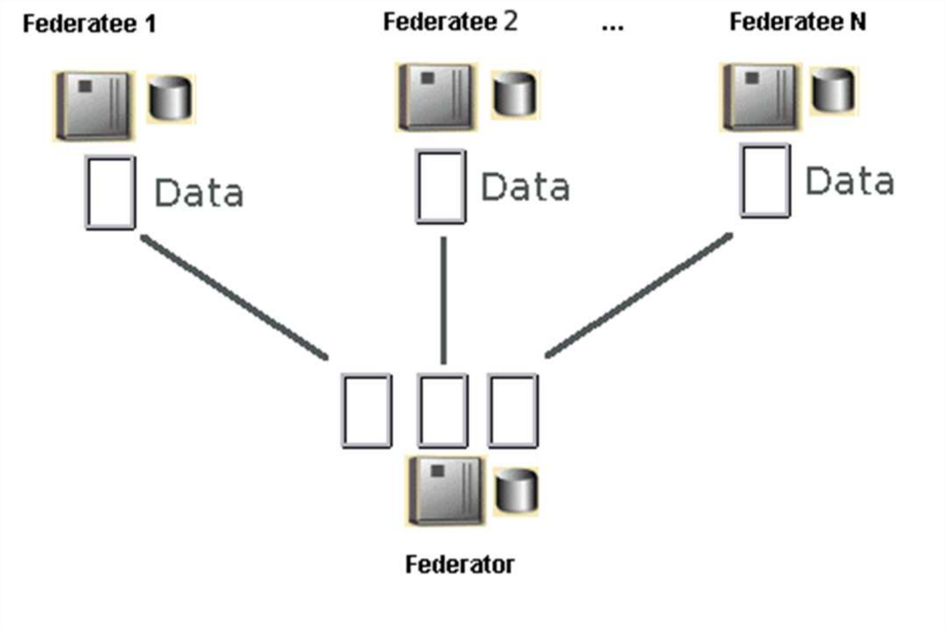
"YES" ... ... ☒

"NO" ... ... □

?

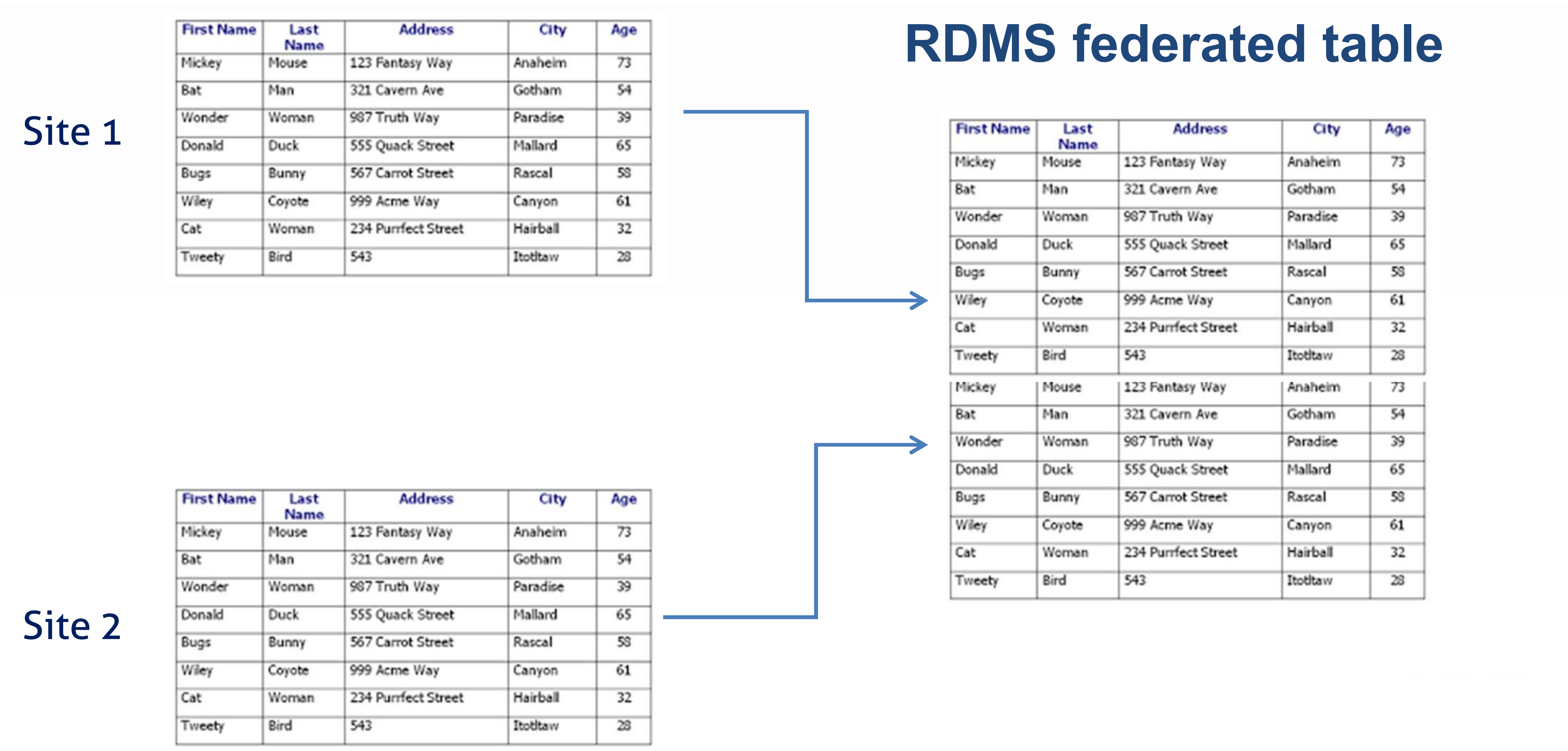


What is Database Federation?



Tools that transparently integrates multiple autonomous databases into a single virtual entity

Federation within RDMS



Site 1

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

Site 2

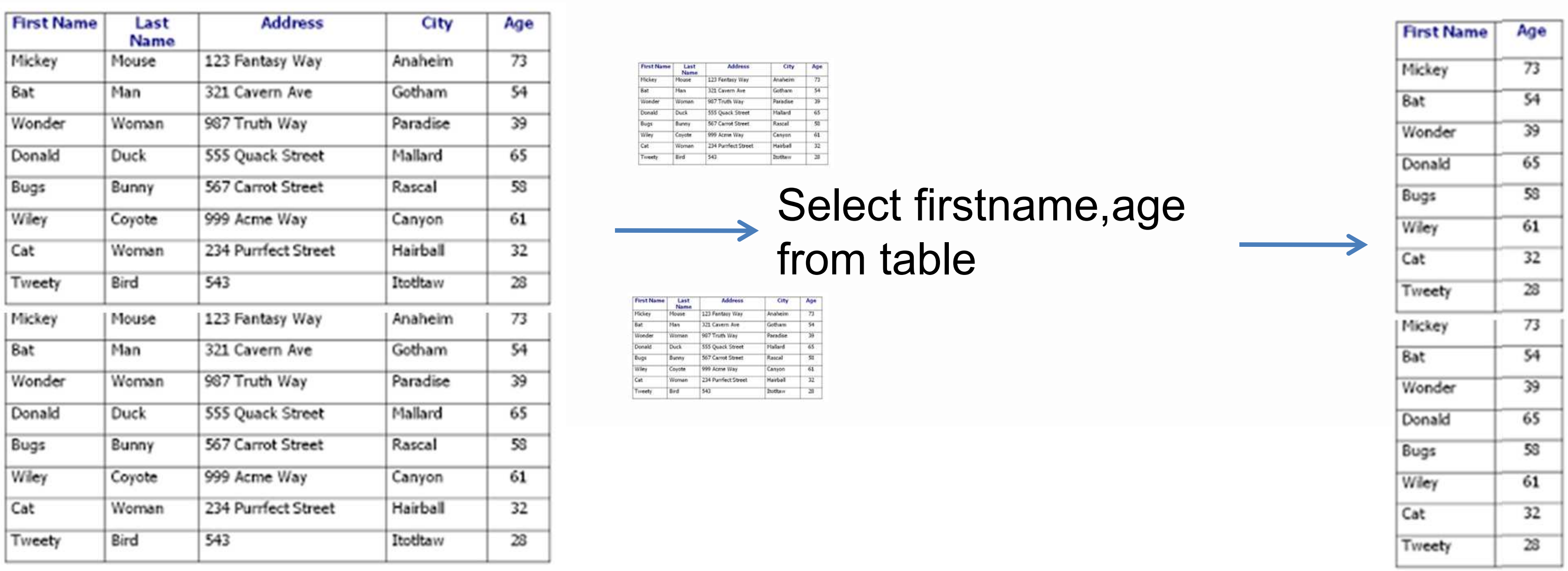
First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

RDMS federated table

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

Federation within RDMS (cont)

RDMS federated table



First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

Select firstname,age from table

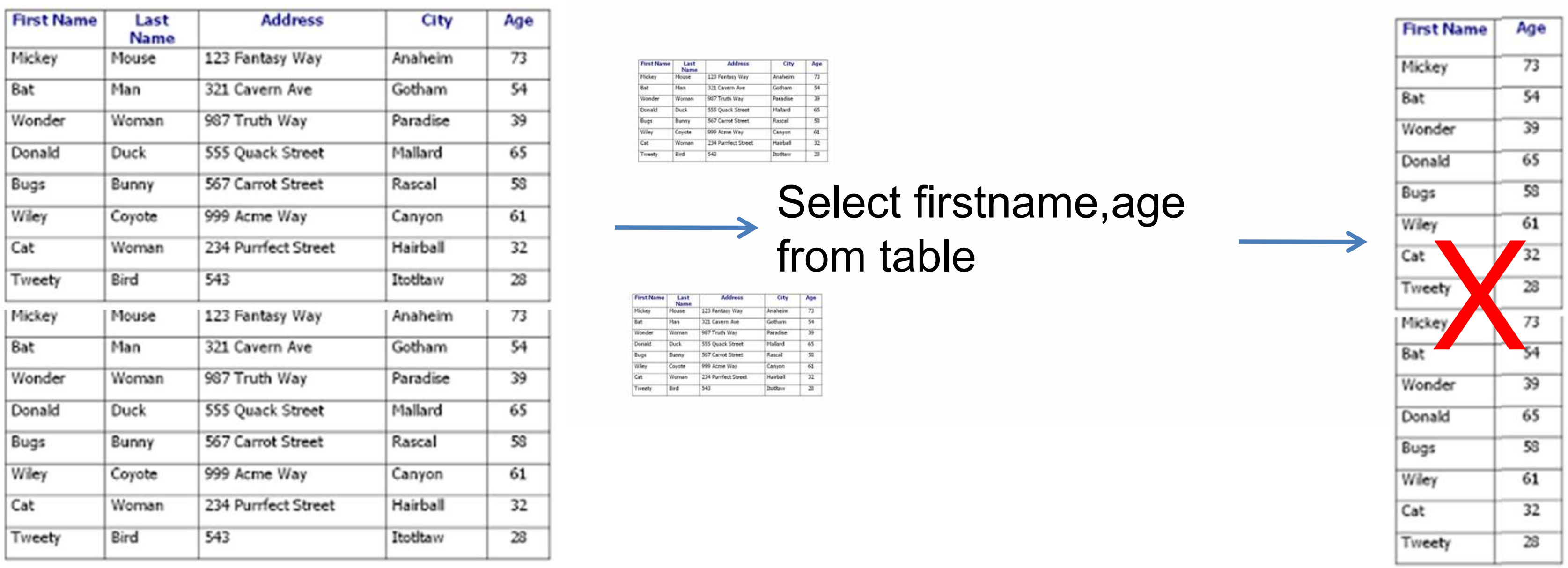
First Name	Age
Mickey	73
Bat	54
Wonder	39
Donald	65
Bugs	58
Wiley	61
Cat	32
Tweety	28
Mickey	73
Bat	54
Wonder	39
Donald	65
Bugs	58
Wiley	61
Cat	32
Tweety	28

Federation within RDMS (cont)

RDMS federated table

PostgreSQL

MySQL



First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

Select firstname,age from table

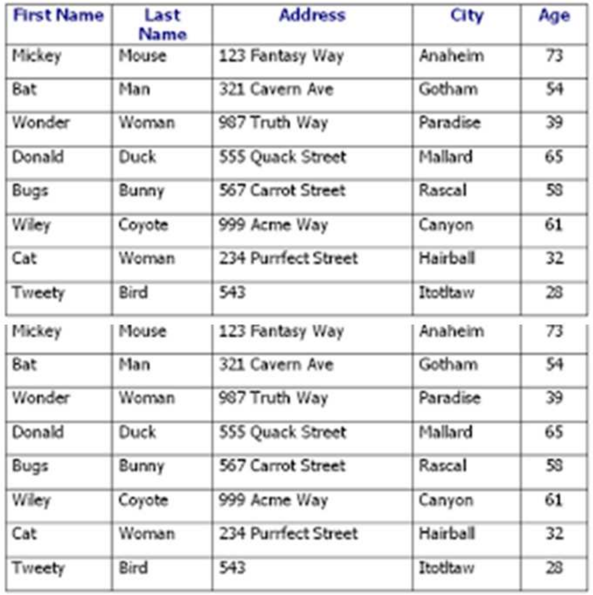
First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

First Name	Age
Mickey	73
Bat	54
Wonder	39
Donald	65
Bugs	58
Wiley	61
Cat	32
Tweety	28
Mickey	73
Bat	54
Wonder	39
Donald	65
Bugs	58
Wiley	61
Cat	32
Tweety	28

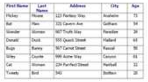
Federation within RDMS (cont)

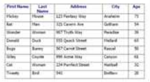
Select firstname,age
from table

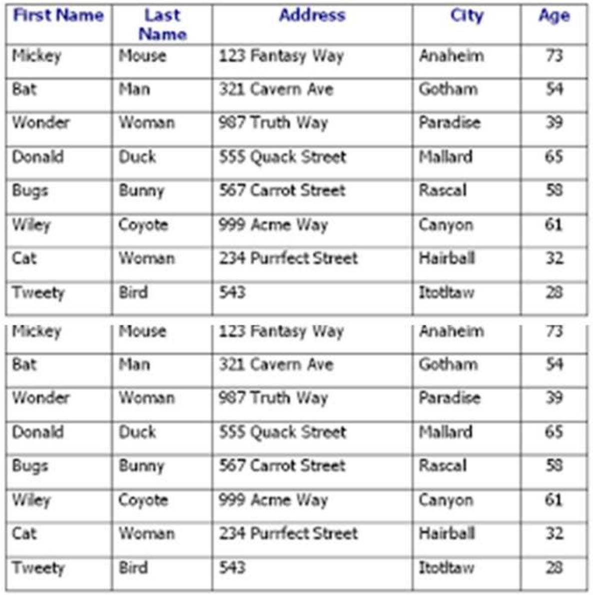
RDMS federated table



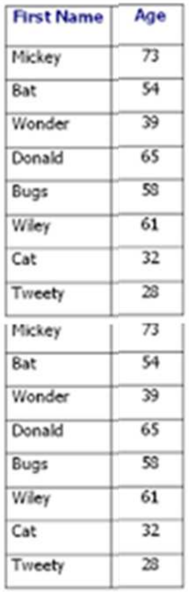
First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28







First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

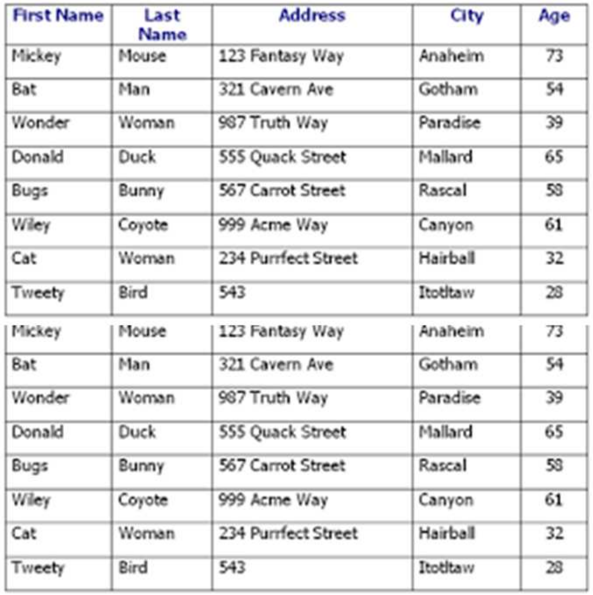


First Name	Age
Mickey	73
Bat	54
Wonder	39
Donald	65
Bugs	58
Wiley	61
Cat	32
Tweety	28
Mickey	73
Bat	54
Wonder	39
Donald	65
Bugs	58
Wiley	61
Cat	32
Tweety	28

Intermediate copy of entire database tables made on the fly at the central site

Federation within RDMS (cont)

RDMS federated table

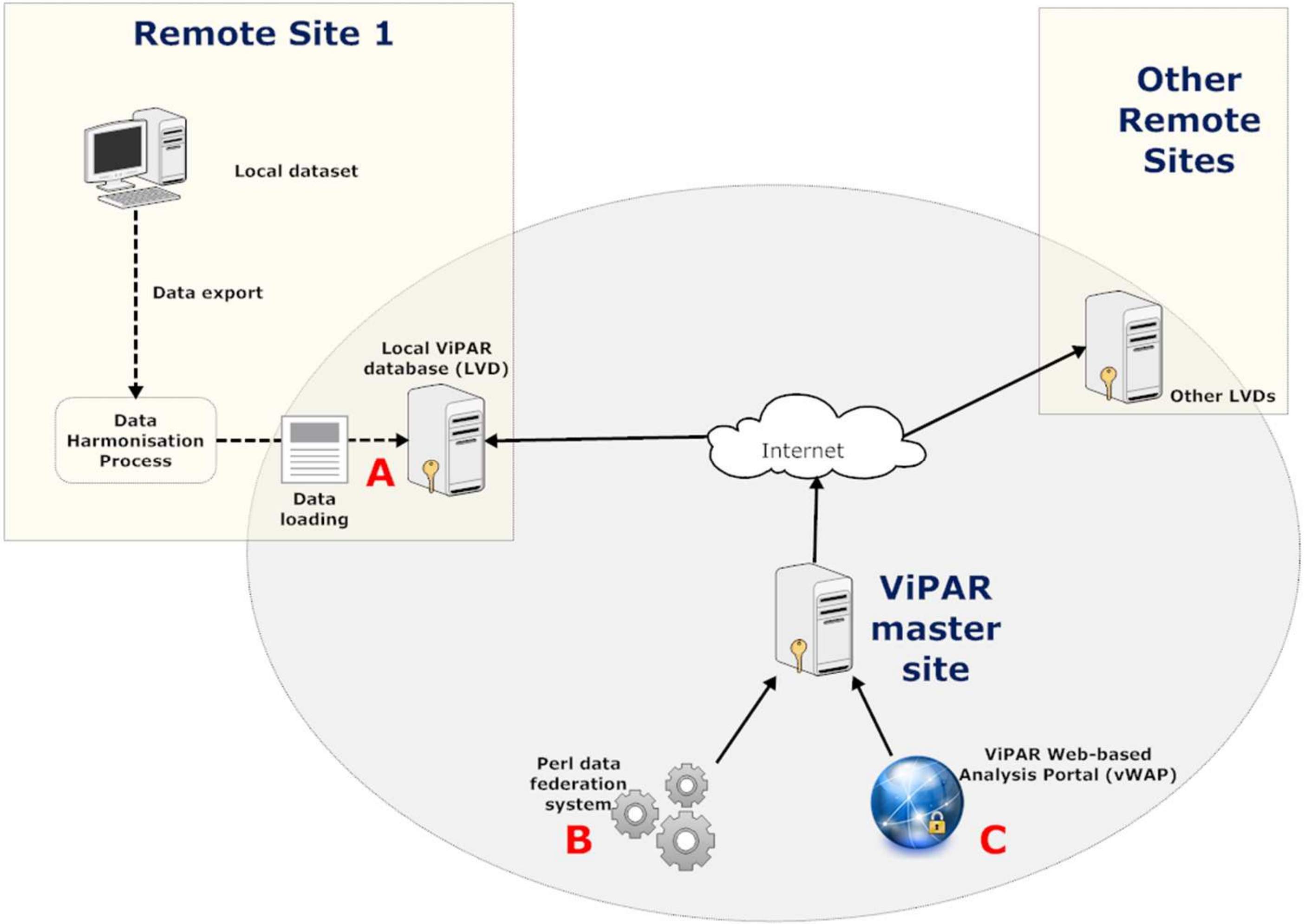


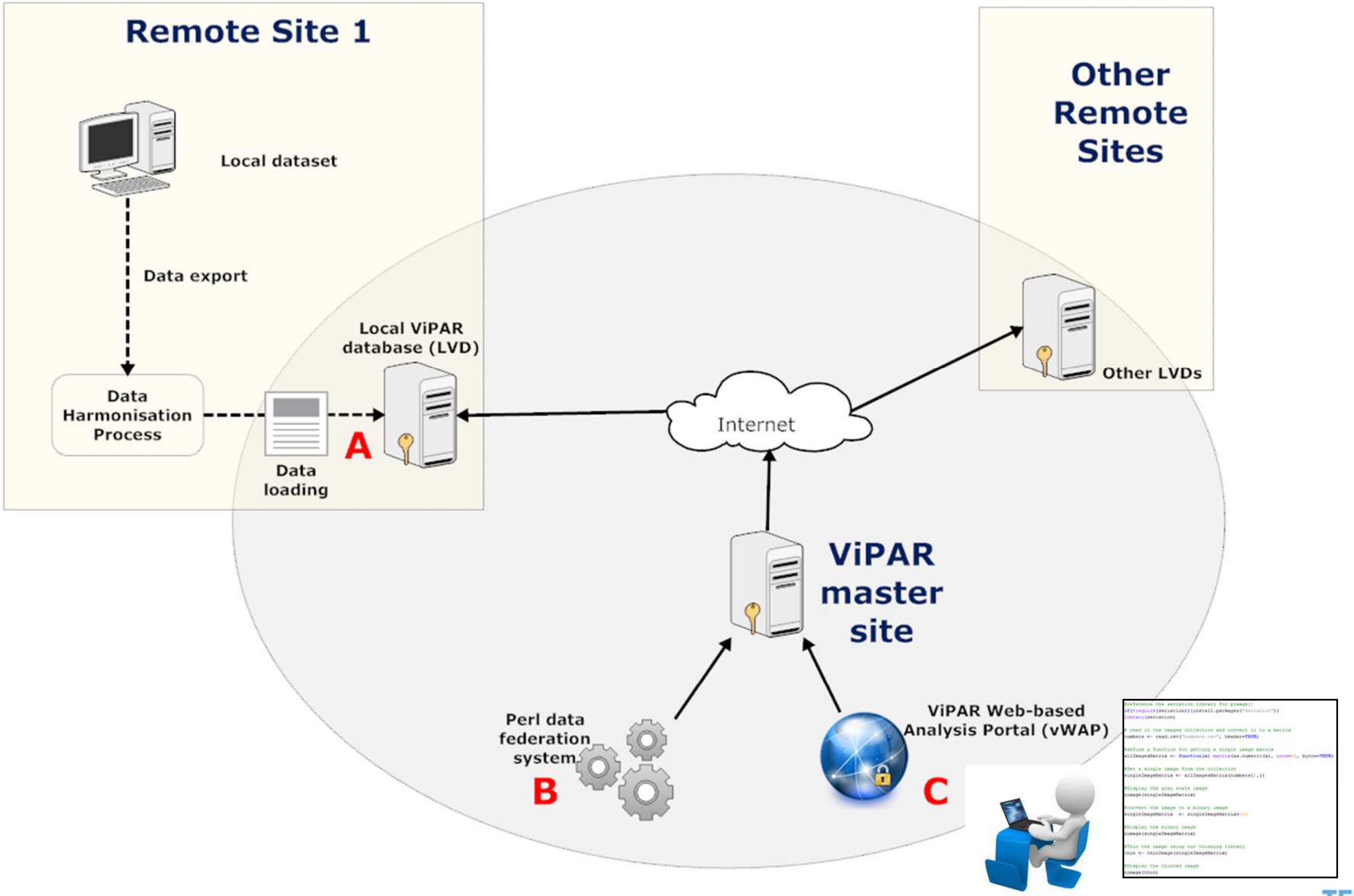
First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

Federation engines in MySQL and Postgres appear to be poorly implemented:
* suitable for LAN only
* suitable for smallish datasets

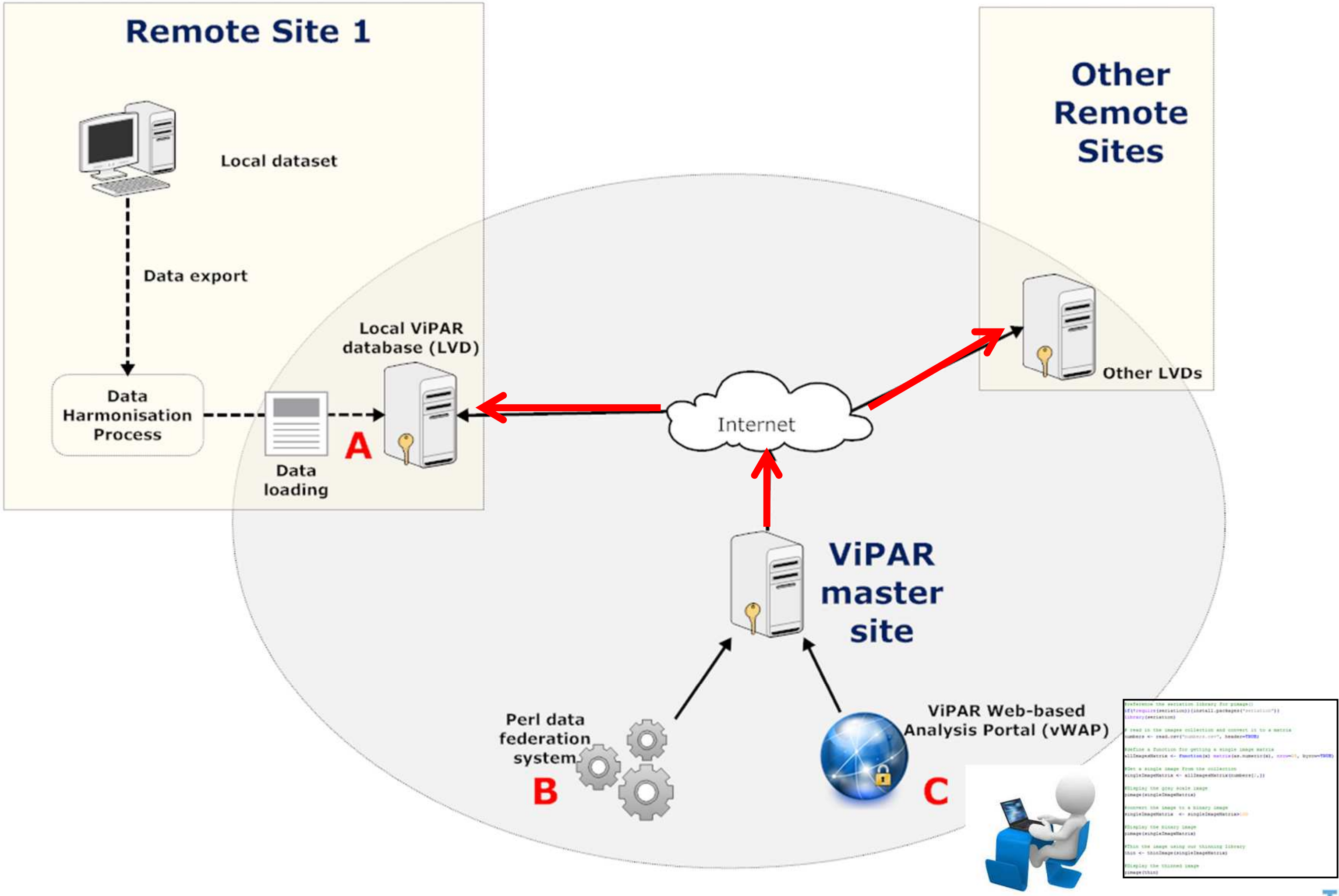
Problem: needs to work securely over Internet with large datasets

ViPAR: overview

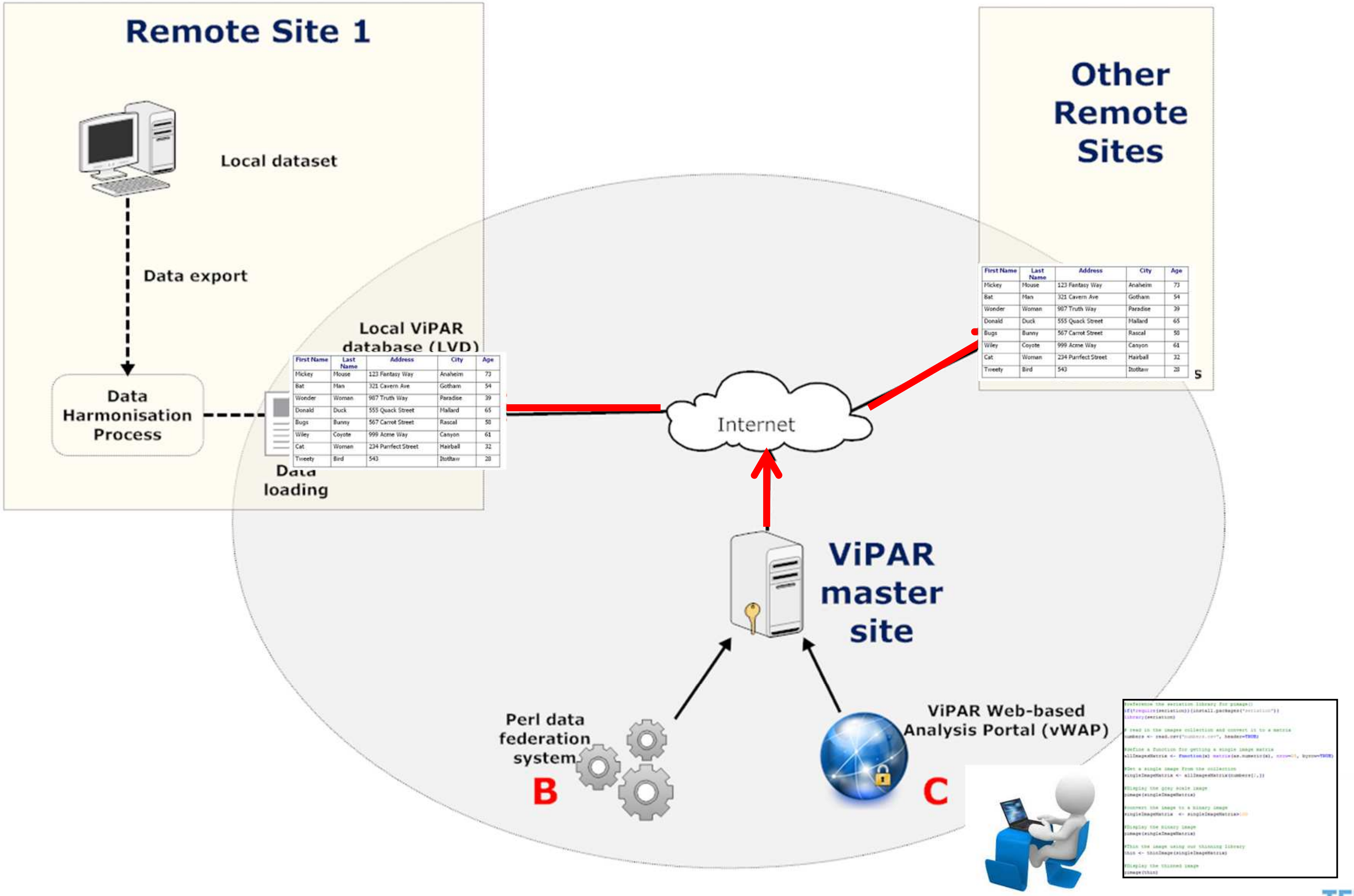

Remote Site 1
Local dataset
Data export
Data Harmonisation Process
Data loading
A
Local ViPAR database (LVD)
Other Remote Sites
Other LVDs
Internet
ViPAR master site
Perl data federation system
B
ViPAR Web-based Analysis Portal (vWAP)
C


Remote Site 1
Local dataset
Data export
Data Harmonisation Process
Data loading
A
Local ViPAR database (LVD)
Internet
Other Remote Sites
Other LVDs
ViPAR master site
Perl data federation system
B
ViPAR Web-based Analysis Portal (vWAP)
C

```
#reference the seriation library for pimage()
if(!require(seriation)){install.packages("seriation")}
library(seriation)

# read in the images collection and convert it to a matrix
numbers <- read.csv("numbers.csv", header=TRUE)

#define a function for getting a single image matrix
allImagesMatrix <- function(x) matrix(as.numeric(x), nrow=[illegible], byrow=TRUE)

#Get a single image from the collection
singleImageMatrix <- allImagesMatrix(numbers[7,])

#Display the grey scale image
pimage(singleImageMatrix)

#convert the image to a binary image
singleImageMatrix  <- singleImageMatrix>[illegible]

#Display the binary image
pimage(singleImageMatrix)

#Thin the image using our thinning library
thin <- thinImage(singleImageMatrix)

#Display the thinned image
pimage(thin)
```


Remote Site 1
Local dataset
Data export
Data Harmonisation Process
Data loading
Local ViPAR database (LVD)
A
Internet
Other Remote Sites
Other LVDs
ViPAR master site
Perl data federation system
B
ViPAR Web-based Analysis Portal (vWAP)
C

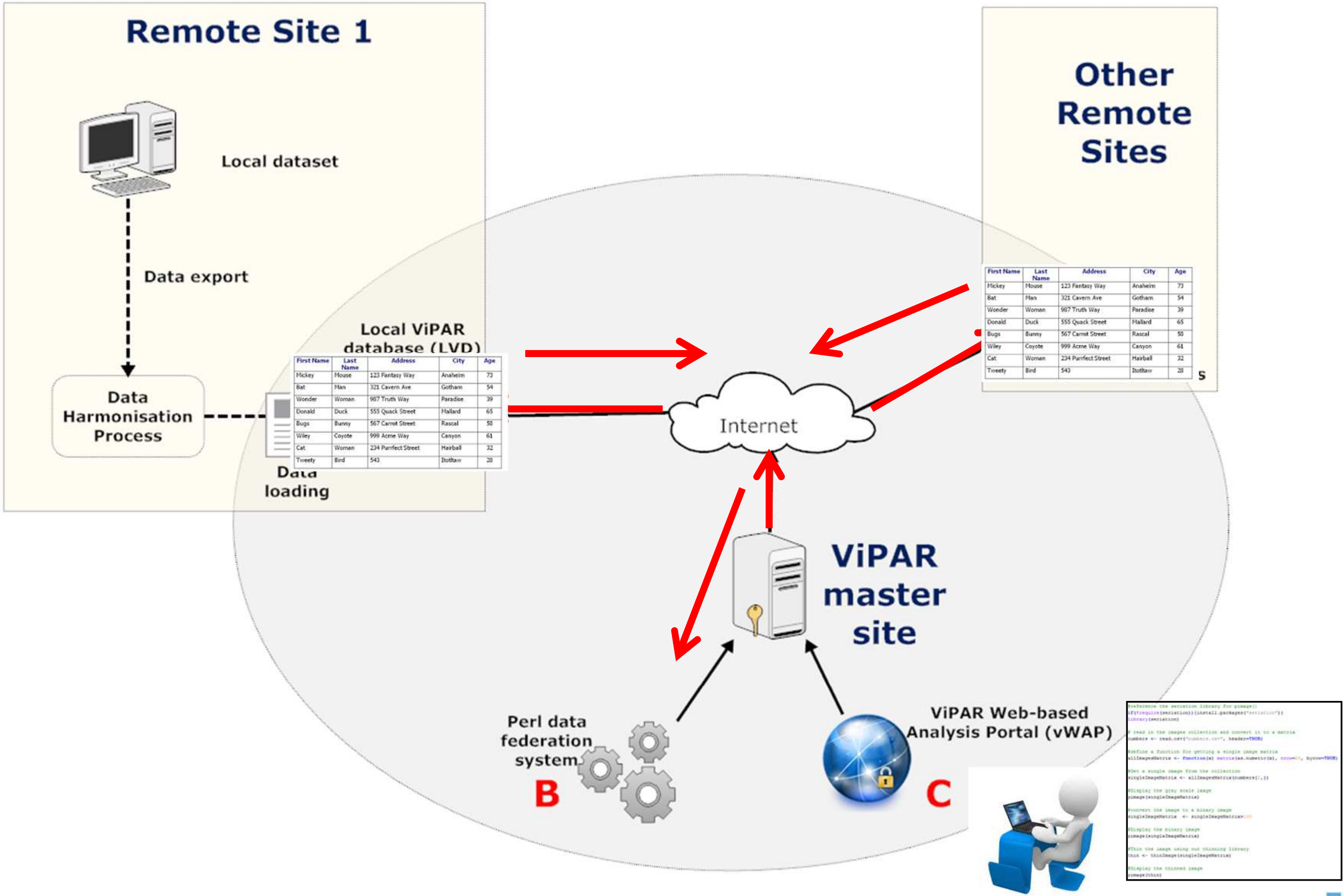
```
#reference the seriation library for pimage()
if(!require(seriation)){install.packages("seriation")}
library(seriation)

# read in the images collection and convert it to a matrix
numbers <- read.csv("numbers.csv", header=TRUE)

#define a function for getting a single image matrix
allImagesMatrix <- function(x) matrix(as.numeric(x), nrow=[illegible], byrow=TRUE)

#Get a single image from the collection
singleImageMatrix <- allImagesMatrix(numbers[7,])

#Display the grey scale image
pimage(singleImageMatrix)

#convert the image to a binary image
singleImageMatrix  <- singleImageMatrix>[illegible]

#Display the binary image
pimage(singleImageMatrix)

#Thin the image using our thinning library
thin <- thinImage(singleImageMatrix)

#Display the thinned image
pimage(thin)
```


Remote Site 1
Local dataset
Data export
Data Harmonisation Process
Local ViPAR database (LVD)
Data loading
Internet
Other Remote Sites
ViPAR master site
Perl data federation system
B
ViPAR Web-based Analysis Portal (vWAP)
C

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

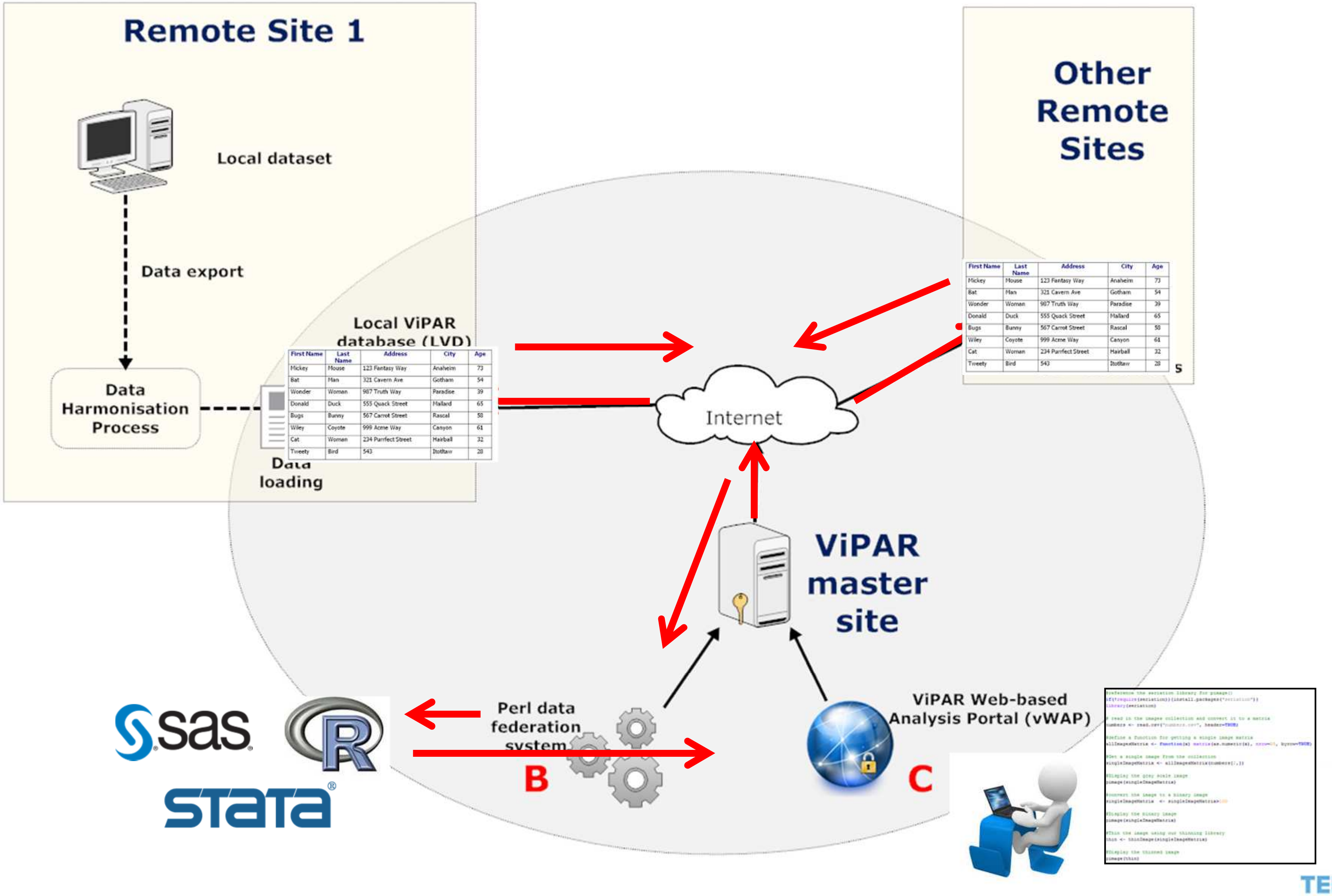
```
#reference the seriation library for pimage()
if(!require(seriation)){install.packages("seriation")}
library(seriation)

# read in the images collection and convert it to a matrix
numbers <- read.csv("numbers.csv", header=TRUE)

#define a function for getting a single image matrix
allImagesMatrix <- function(x) matrix(as.numeric(x), nrow=[illegible], byrow=TRUE)

#Get a single image from the collection
singleImageMatrix <- allImagesMatrix(numbers[7,])

#Display the grey scale image
pimage(singleImageMatrix)

#convert the image to a binary image
singleImageMatrix  <- singleImageMatrix>[illegible]

#Display the binary image
pimage(singleImageMatrix)

#Thin the image using our thinning library
thin <- thinImage(singleImageMatrix)

#Display the thinned image
pimage(thin)
```


Remote Site 1
Local dataset
Data export
Data Harmonisation Process
Local ViPAR database (LVD)
Data loading
Other Remote Sites
s
Internet
ViPAR master site
Perl data federation system
B
ViPAR Web-based Analysis Portal (vWAP)
C

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

```
#reference the seriation library for pimage()
if(!require(seriation)){install.packages("seriation")}
library(seriation)

# read in the images collection and convert it to a matrix
numbers <- read.csv("numbers.csv", header=TRUE)

#define a function for getting a single image matrix
allImagesMatrix <- function(x) matrix(as.numeric(x), nrow=[illegible], byrow=TRUE)

#Get a single image from the collection
singleImageMatrix <- allImagesMatrix(numbers[7,])

#Display the grey scale image
pimage(singleImageMatrix)

#convert the image to a binary image
singleImageMatrix  <- singleImageMatrix>[illegible]

#Display the binary image
pimage(singleImageMatrix)

#Thin the image using our thinning library
thin <- thinImage(singleImageMatrix)

#Display the thinned image
pimage(thin)
```



Remote Site 1
Local dataset
Data export
Data Harmonisation Process
Local ViPAR database (LVD)
Data loading
Other Remote Sites
s
Internet
ViPAR master site
Perl data federation system
B
ViPAR Web-based Analysis Portal (vWAP)
C
sas
R
STATA

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

```
#reference the seriation library for pimage()
if(!require(seriation)){install.packages("seriation")}
library(seriation)

# read in the images collection and convert it to a matrix
numbers <- read.csv("numbers.csv", header=TRUE)

#define a function for getting a single image matrix
allImagesMatrix <- function(x) matrix(as.numeric(x), nrow=[illegible], byrow=TRUE)

#Get a single image from the collection
singleImageMatrix <- allImagesMatrix(numbers[[illegible],])

#Display the grey scale image
pimage(singleImageMatrix)

#convert the image to a binary image
singleImageMatrix  <- singleImageMatrix>[illegible]

#Display the binary image
pimage(singleImageMatrix)

#Thin the image using our thinning library
thin <- thinImage(singleImageMatrix)

#Display the thinned image
pimage(thin)
```

ViPAR – a custom data federation platform

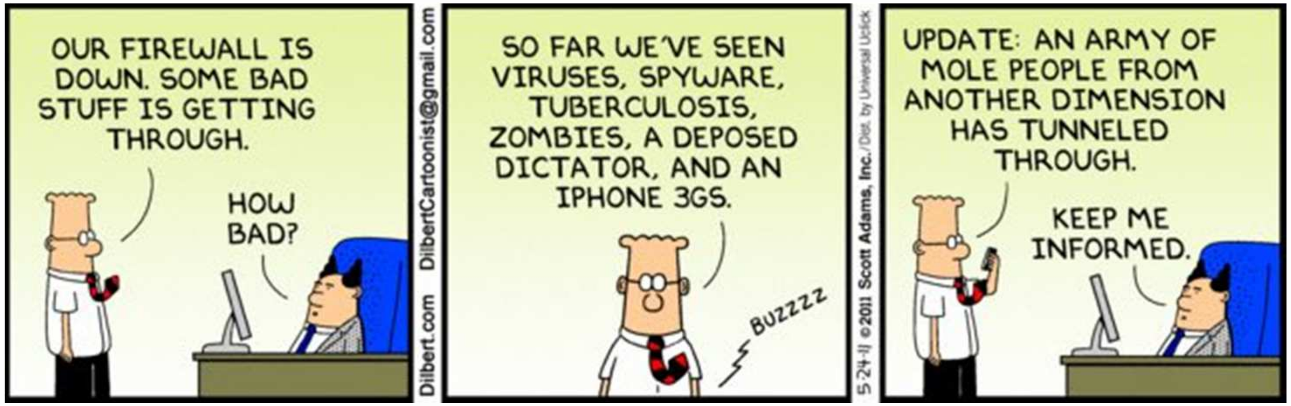
Open-source stack – Linux, Apache, Mysql, OpenSSL
Available as a pre-built VM images for easy access

http://bioinformatics.childhealthresearch.org.au/software/vipar/
(pre-built VM images for easy use – Vbox and Vmware)
https://gitlab.com/kim.carter/ViPAR (source code)
https://groups.google.com/forum/#!forum/vipar (support forum)

ViPAR Daemon – 1500 lines of Perl code
-Handles multiple clients simultaneously, and can perform eg parallel data retrieval (all sites at the same time)
-Implements its own federation (and analysis) engine

VWAP – (ViPAR Web-based Analysis Portal)
-8000 lines of Perl code
-1200 lines of Javascript

Behind the scenes


OUR FIREWALL IS DOWN. SOME BAD STUFF IS GETTING THROUGH.
HOW BAD?
SO FAR WE'VE SEEN VIRUSES, SPYWARE, TUBERCULOSIS, ZOMBIES, A DEPOSED DICTATOR, AND AN IPHONE 3GS.
BUZZZZ
UPDATE: AN ARMY OF MOLE PEOPLE FROM ANOTHER DIMENSION HAS TUNNELED THROUGH.
KEEP ME INFORMED.
Dilbert.com DilbertCartoonist@gmail.com
5-24-11 ©2011 Scott Adams, Inc./Dist. by Universal Uclick

Behind the scenes

MySQL

Open-source database management system, with powerful features including database federation

Widely used, especially for large databases

SSH

Powerful, free solution to securely exchange data between 2 networked devices over an insecure channel (such as the Internet)

Enterprise reliability and security

“tunnels” allow us to transfer of programs services between remote computers dynamically

SSL

Client-side and Server-side SSL certificates used (along with user accounts and roles) to provide layered, strong security model

Data model

Study
Overarching conglomeration of analysis projects, data resources, data dictionary and variables and users

Data dictionary
A version controlled set of Variables and Missing data fields

Variable
A data field of a given type (eg continuous, categorical, date) with a missing data type

Sites, Servers and Resources
A site is a nominal geo-location housing a ViPAR LVD server accessible on a specified SSH port, with one or more (certified) harmonised dataset resources on the server

Data model (cont)

Analysis “projects”

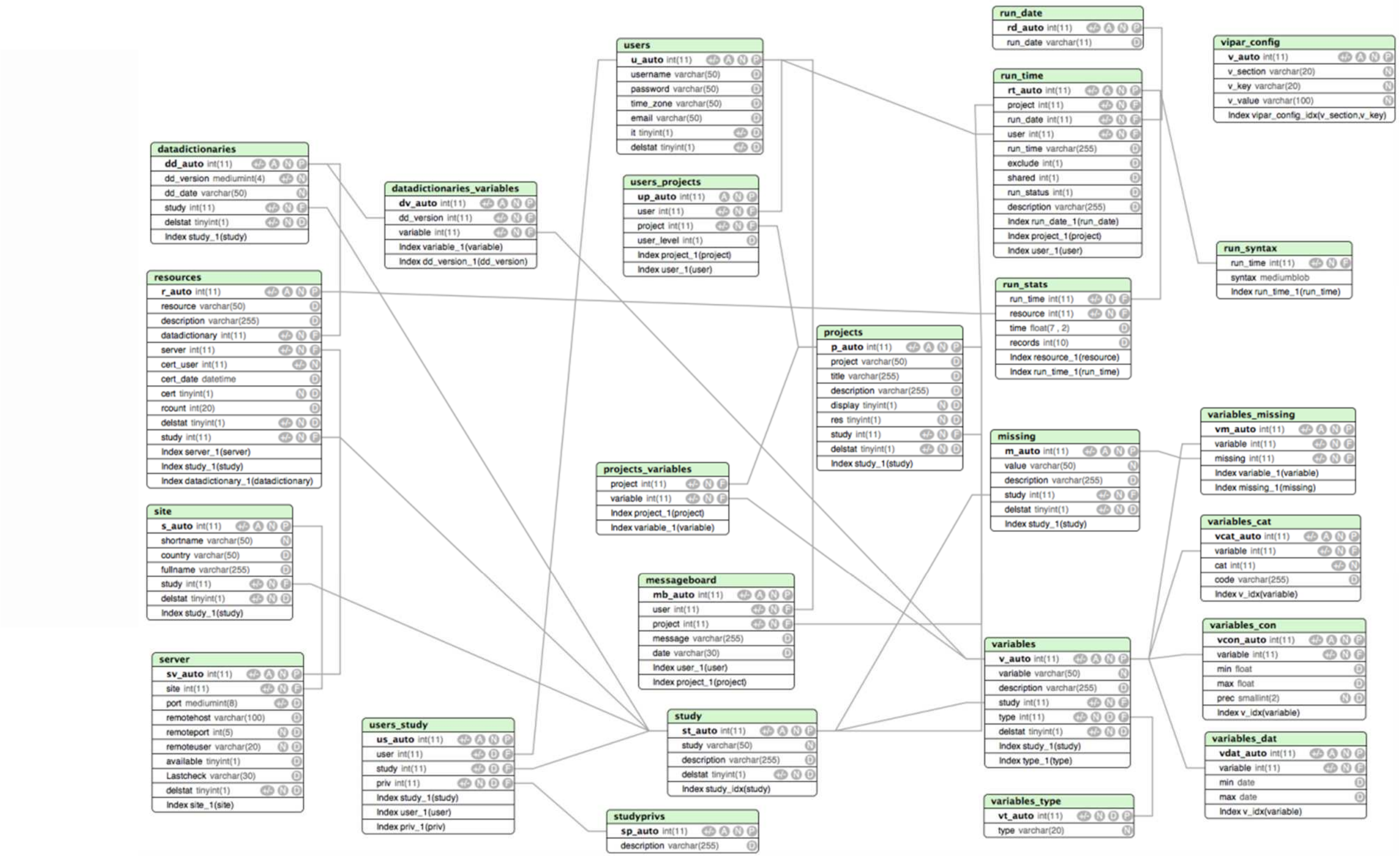
Defined research question(s) (eg agreed by AOC)

- Specified list of Resources to be included
- Specified list of Variables available for analysis
- Specified list of Users with access @ analyst or guest level (read-only)
- Results/outputs can be tagged as “sharing” accessible to enable non-project users within ViPAR to see specific results

Users

- Can be designated as “Admin” (see all projects and get extra admin menu options”
- Can be member of multiple analysis projects
- Unique user/pass, along with client SSL certificate (recommended)

Behind the scenes: data model



datadictionaries	
dd_auto int(11)	</> A N P
dd_version mediumint(4)	</> N
dd_date varchar(50)	N
study int(11)	</> N F
delstat tinyint(1)	</> N D
Index study_1(study)	

datadictionaries_variables	
dv_auto int(11)	</> A N P
dd_version int(11)	</> N F
variable int(11)	</> N F
Index variable_1(variable)	
Index dd_version_1(dd_version)	

resources	
r_auto int(11)	</> A N P
resource varchar(50)	D
description varchar(255)	D
datadictionary int(11)	</> N F
server int(11)	</> N F
cert_user int(11)	</> N
cert_date datetime	D
cert tinyint(1)	N D
rcount int(20)	D
delstat tinyint(1)	</> N D
study int(11)	</> N F
Index server_1(server)	
Index study_1(study)	
Index datadictionary_1(datadictionary)	

site	
s_auto int(11)	</> A N P
shortname varchar(50)	N
country varchar(50)	D
fullname varchar(255)	D
study int(11)	</> N F
delstat tinyint(1)	</> N D
Index study_1(study)	

server	
sv_auto int(11)	</> A N P
site int(11)	</> N F
port mediumint(8)	</> D
remotehost varchar(100)	D
remoteport int(5)	N D
remoteuser varchar(20)	N D
available tinyint(1)	D
Lastcheck varchar(30)	D
delstat tinyint(1)	</> N D
Index site_1(site)	

users_study	
us_auto int(11)	</> A N P
user int(11)	</> D F
study int(11)	</> D F
priv int(11)	</> N D F
Index study_1(study)	
Index user_1(user)	
Index priv_1(priv)	

users	
u_auto int(11)	</> A N P
username varchar(50)	D
password varchar(50)	D
time_zone varchar(50)	D
email varchar(50)	D
it tinyint(1)	</> D
delstat tinyint(1)	</> D

users_projects	
up_auto int(11)	A N P
user int(11)	</> N F
project int(11)	</> N F
user_level int(1)	D
Index project_1(project)	
Index user_1(user)	

projects_variables	
project int(11)	</> N F
variable int(11)	</> N F
Index project_1(project)	
Index variable_1(variable)	

messageboard	
mb_auto int(11)	</> A N P
user int(11)	</> N F
project int(11)	</> N F
message varchar(255)	D
date varchar(30)	D
Index user_1(user)	
Index project_1(project)	

study	
st_auto int(11)	</> A N P
study varchar(50)	N
description varchar(255)	D
delstat tinyint(1)	</> N D
Index study_idx(study)	

studyprivs	
sp_auto int(11)	</> A N P
description varchar(255)	D

projects	
p_auto int(11)	</> A N P
project varchar(50)	D
title varchar(255)	D
description varchar(255)	D
display tinyint(1)	N D
res tinyint(1)	N D
study int(11)	</> N F
delstat tinyint(1)	</> N D
Index study_1(study)	

run_date	
rd_auto int(11)	</> A N P
run_date varchar(11)	D

run_time	
rt_auto int(11)	</> A N P
project int(11)	</> N F
run_date int(11)	</> N F
user int(11)	</> N F
run_time varchar(255)	D
exclude int(1)	D
shared int(1)	D
run_status int(1)	D
description varchar(255)	D
Index run_date_1(run_date)	
Index project_1(project)	
Index user_1(user)	

run_stats	
run_time int(11)	</> N F
resource int(11)	</> N F
time float(7 , 2)	D
records int(10)	D
Index resource_1(resource)	
Index run_time_1(run_time)	

missing	
m_auto int(11)	</> A N P
value varchar(50)	N
description varchar(255)	D
study int(11)	</> N F
delstat tinyint(1)	</> N D
Index study_1(study)	

variables	
v_auto int(11)	</> A N P
variable varchar(50)	N
description varchar(255)	D
study int(11)	</> N F
type int(11)	</> N D F
delstat tinyint(1)	</> N D
Index study_1(study)	
Index type_1(type)	

variables_type	
vt_auto int(11)	</> N D P
type varchar(20)	N

vipar_config	
v_auto int(11)	</> A N P
v_section varchar(20)	N
v_key varchar(20)	N
v_value varchar(100)	N
Index vipar_config_idx(v_section,v_key)	

run_syntax	
run_time int(11)	</> N F
syntax mediumblob	
Index run_time_1(run_time)	

variables_missing	
vm_auto int(11)	</> A N P
variable int(11)	</> N F
missing int(11)	</> N F
Index variable_1(variable)	
Index missing_1(missing)	

variables_cat	
vcat_auto int(11)	</> A N P
variable int(11)	</> N F
cat int(11)	</> N
code varchar(255)	D
Index v_idx(variable)	

variables_con	
vcon_auto int(11)	</> A N P
variable int(11)	</> N F
min float	D
max float	D
prec smallint(2)	N D
Index v_idx(variable)	

variables_dat	
vdat_auto int(11)	</> A N P
variable int(11)	</> N F
min date	D
max date	D
Index v_idx(variable)	

TELETHON KIDS INSTITUTE
Discover. Prevent. Cure.

How does ViPAR protect my data?

Network level

- Private, encrypted network connecting each data site (LVD) to the master site (VMS). Protected by hardware & software firewalls
- Only the VMS can initiate connections to the LVDs
- Sites maintain their own data (in the LVD) – no data is saved at the VMS

User-level

- Unique user pass & cert combo (with client SSL certificates) – you can't get to the site login page without
- Only users who run code get to "pool" the data (within a defined project)
- Implicit "trust" to give max flexibility in analysis -> that users won't try to access individual data

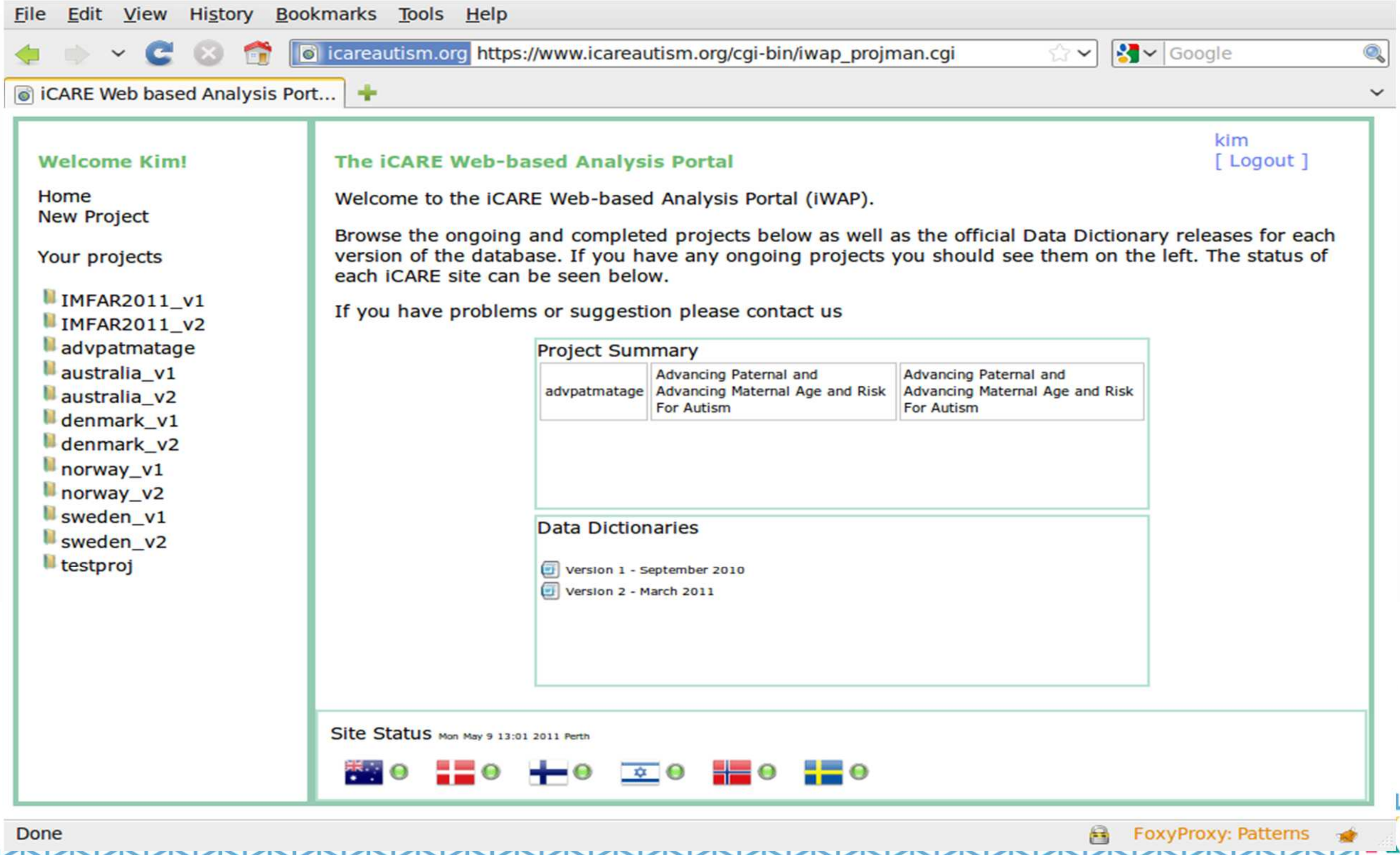
Portal-level

- Everything is logged, all run code, all interactions including "deletions".
- No direct access to the data

Data-level design

- Design your dictionary to minimise potentially identifying fields

ViPAR Portal (for iCARE)



File Edit View History Bookmarks Tools Help
icareautism.org https://www.icareautism.org/cgi-bin/iwap_projman.cgi
Google
iCARE Web based Analysis Port...

Welcome Kim!

Home
New Project

Your projects

- IMFAR2011_v1
- IMFAR2011_v2
- advpatmatage
- australia_v1
- australia_v2
- denmark_v1
- denmark_v2
- norway_v1
- norway_v2
- sweden_v1
- sweden_v2
- testproj

kim
[Logout]

The iCARE Web-based Analysis Portal

Welcome to the iCARE Web-based Analysis Portal (iWAP).

Browse the ongoing and completed projects below as well as the official Data Dictionary releases for each version of the database. If you have any ongoing projects you should see them on the left. The status of each iCARE site can be seen below.

If you have problems or suggestion please contact us

Project Summary

advpatmatage	Advancing Paternal and Advancing Maternal Age and Risk For Autism	Advancing Paternal and Advancing Maternal Age and Risk For Autism

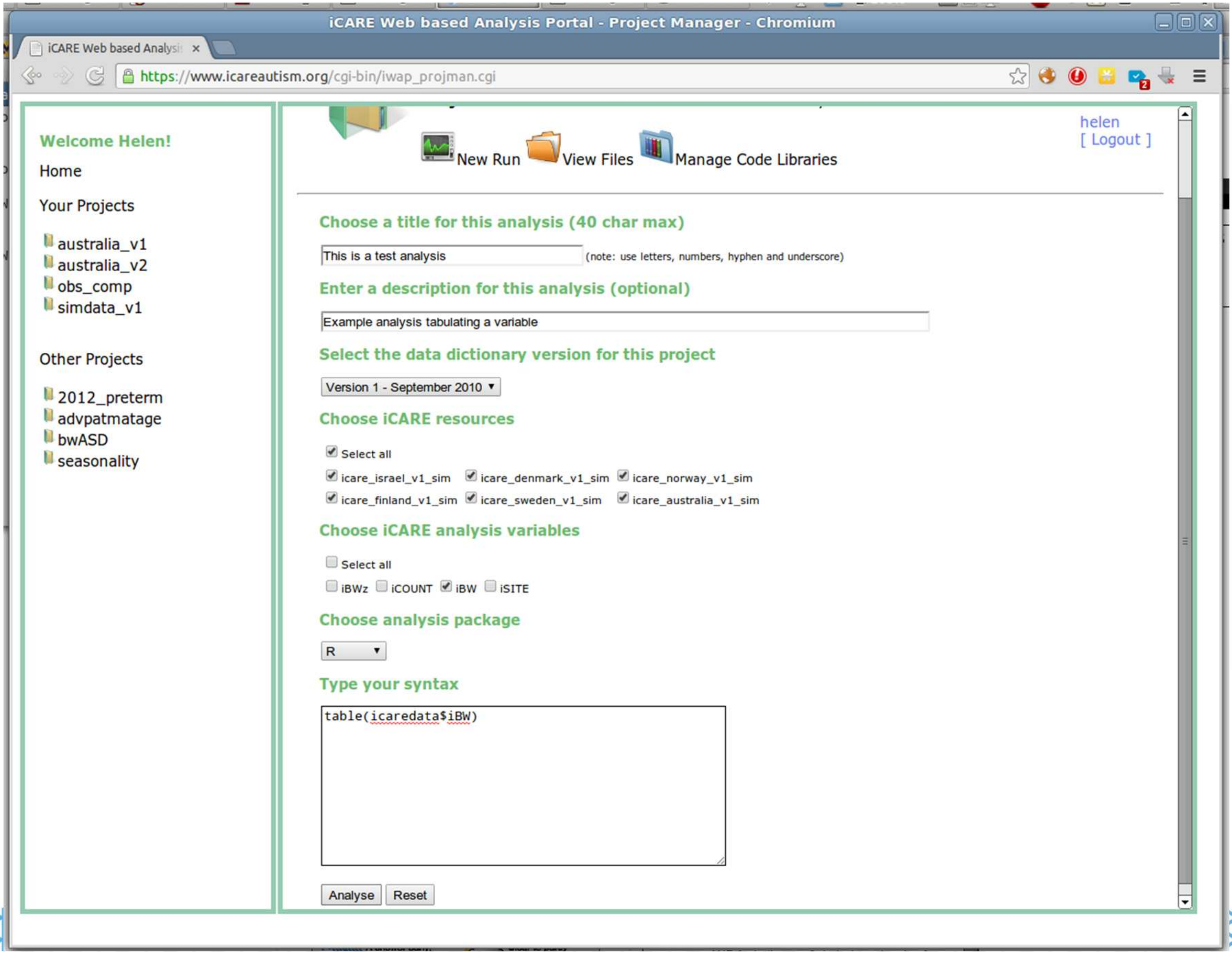
Data Dictionaries

- Version 1 - September 2010
- Version 2 - March 2011

Site Status Mon May 9 13:01 2011 Perth

● ● ● ● ● ●

Running an analysis


iCARE Web based Analysis Portal - Project Manager - Chromium
iCARE Web based Analysis
https://www.icareautism.org/cgi-bin/iwap_projman.cgi
New Run
View Files
Manage Code Libraries
helen
[Logout]

Welcome Helen!

Home

Your Projects

- australia_v1
- australia_v2
- obs_comp
- simdata_v1

Other Projects

- 2012_preterm
- advpatmatage
- bwASD
- seasonality

Choose a title for this analysis (40 char max)

This is a test analysis (note: use letters, numbers, hyphen and underscore)

Enter a description for this analysis (optional)

Example analysis tabulating a variable

Select the data dictionary version for this project

Version 1 - September 2010 ▾

Choose iCARE resources

☑ Select all

☑ icare_israel_v1_sim ☑ icare_denmark_v1_sim ☑ icare_norway_v1_sim

☑ icare_finland_v1_sim ☑ icare_sweden_v1_sim ☑ icare_australia_v1_sim

Choose iCARE analysis variables

□ Select all

□ iBWz □ iCOUNT ☑ iBW □ iSITE

Choose analysis package

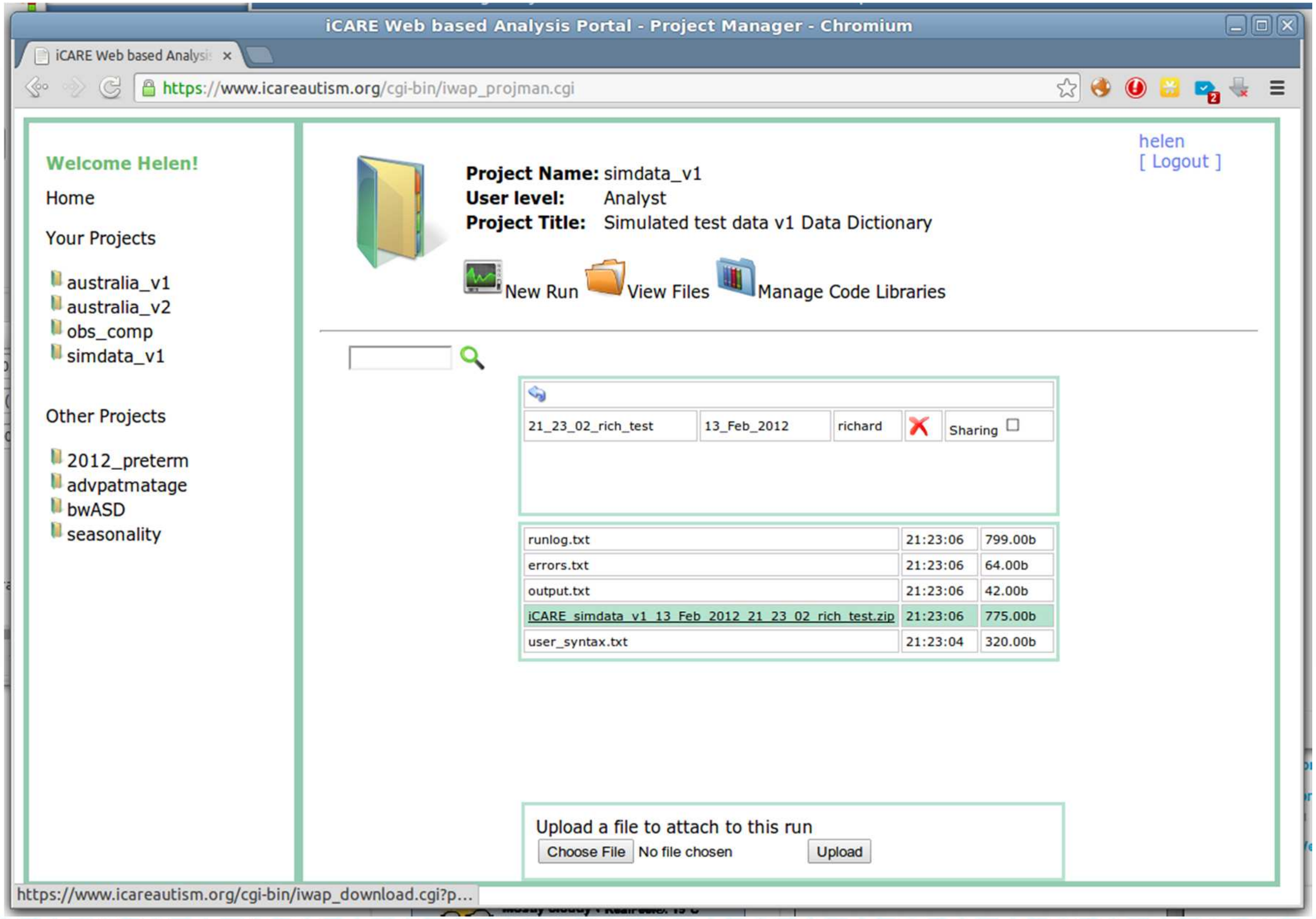
R ▾

Type your syntax

```
table(icaredata$iBW)
```

Analyse Reset

Accessing outputs



iCARE Web based Analysis Portal - Project Manager - Chromium

https://www.icareautism.org/cgi-bin/iwap_projman.cgi

Welcome Helen!

Home

Your Projects

- australia_v1
- australia_v2
- obs_comp
- simdata_v1

Other Projects

- 2012_preterm
- advpatmatage
- bwASD
- seasonality

helen
[Logout]

Project Name: simdata_v1
User level: Analyst
Project Title: Simulated test data v1 Data Dictionary

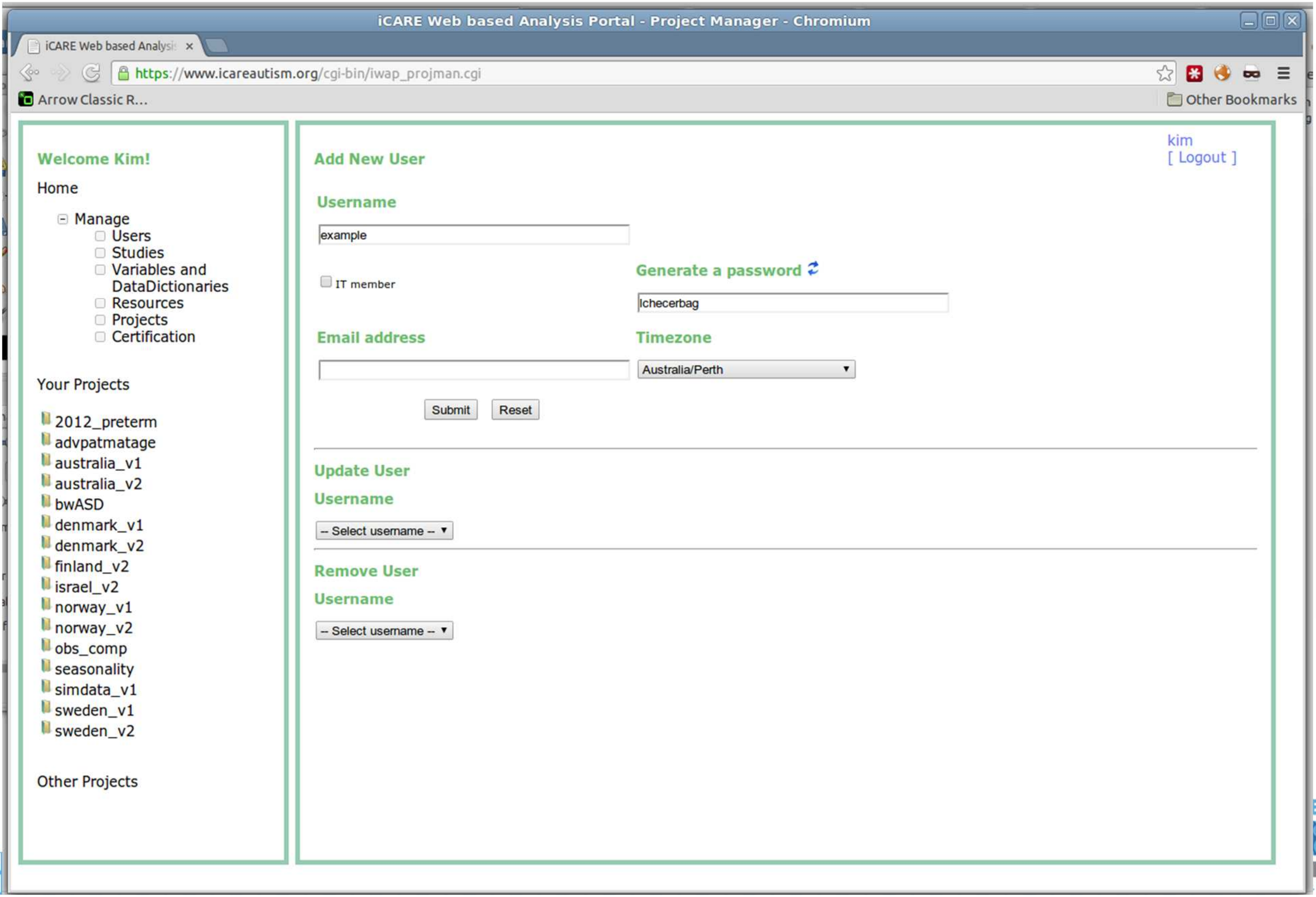
21_23_02_rich_test	13_Feb_2012	richard	✗	Sharing □

runlog.txt	21:23:06	799.00b
errors.txt	21:23:06	64.00b
output.txt	21:23:06	42.00b
iCARE_simdata_v1_13_Feb_2012_21_23_02_rich_test.zip	21:23:06	775.00b
user_syntax.txt	21:23:04	320.00b

Upload a file to attach to this run

https://www.icareautism.org/cgi-bin/iwap_download.cgi?p...

Management



iCARE Web based Analysis Portal - Project Manager - Chromium
iCARE Web based Analysis
https://www.icareautism.org/cgi-bin/iwap_projman.cgi
Arrow Classic R...
Other Bookmarks

Welcome Kim!

Home

- ⊟ Manage
  - □ Users
  - □ Studies
  - □ Variables and DataDictionaries
  - □ Resources
  - □ Projects
  - □ Certification

Your Projects

- 2012_preterm
- advpatmatage
- australia_v1
- australia_v2
- bwASD
- denmark_v1
- denmark_v2
- finland_v2
- israel_v2
- norway_v1
- norway_v2
- obs_comp
- seasonality
- simdata_v1
- sweden_v1
- sweden_v2

Other Projects

kim
[Logout]

Add New User

Username: example

□ IT member

Generate a password: lchecerbag

Email address:

Timezone: Australia/Perth

Update User

Username: -- Select username --

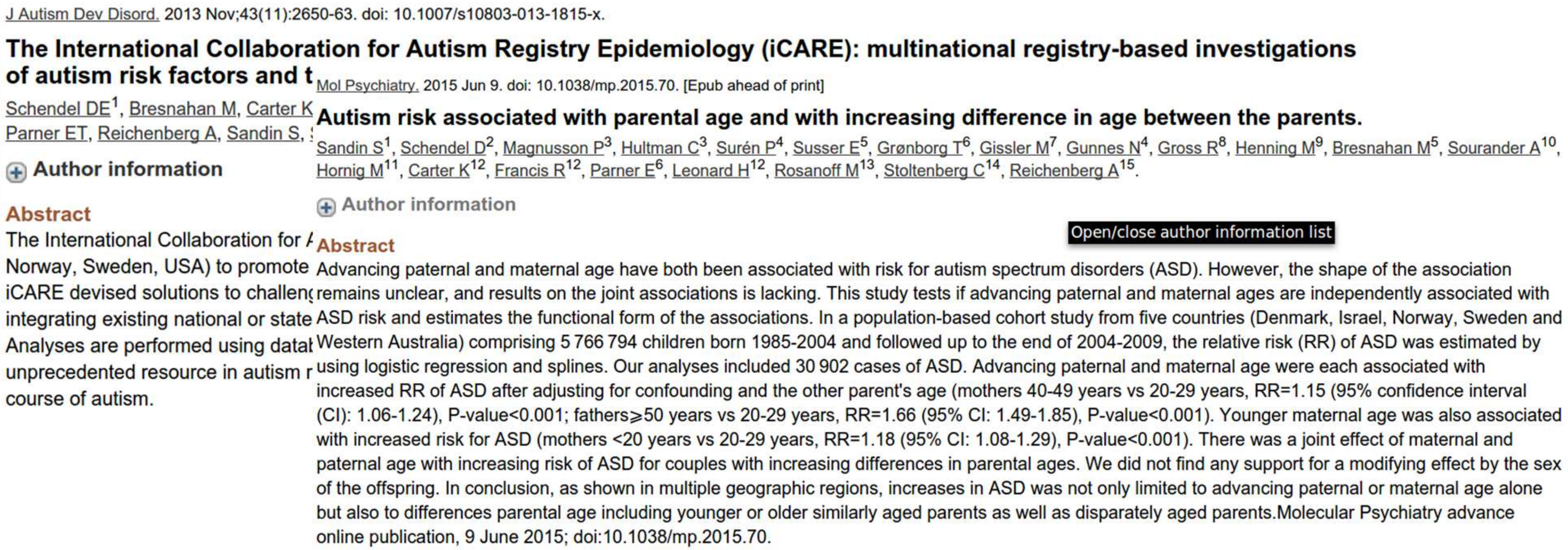
Remove User

Username: -- Select username --

Has ViPAR been successful?

Currently:

- 7 remote sites (Nor,Fin,Den,Swe,Isr,US,WA), containing approx 9million records across 60 harmonised fields
- Passed multiple ethics approvals (at each site)
- Data retrieval with ViPAR: approx 2minutes in serial (one site at a time); approx 30-40 seconds in parallel



J Autism Dev Disord. 2013 Nov;43(11):2650-63. doi: 10.1007/s10803-013-1815-x.

The International Collaboration for Autism Registry Epidemiology (iCARE): multinational registry-based investigations of autism risk factors and t

Schendel DE[1], Bresnahan M, Carter K
Parner ET, Reichenberg A, Sandin S,

⊕ Author information

Abstract

The International Collaboration for A
Norway, Sweden, USA) to promote
iCARE devised solutions to challeng
integrating existing national or state
Analyses are performed using datab
unprecedented resource in autism r
course of autism.

Mol Psychiatry. 2015 Jun 9. doi: 10.1038/mp.2015.70. [Epub ahead of print]

Autism risk associated with parental age and with increasing difference in age between the parents.

Sandin S[1], Schendel D[2], Magnusson P[3], Hultman C[3], Surén P[4], Susser E[5], Grønborg T[6], Gissler M[7], Gunnes N[4], Gross R[8], Henning M[9], Bresnahan M[5], Sourander A[10], Hornig M[11], Carter K[12], Francis R[12], Parner E[6], Leonard H[12], Rosanoff M[13], Stoltenberg C[14], Reichenberg A[15].

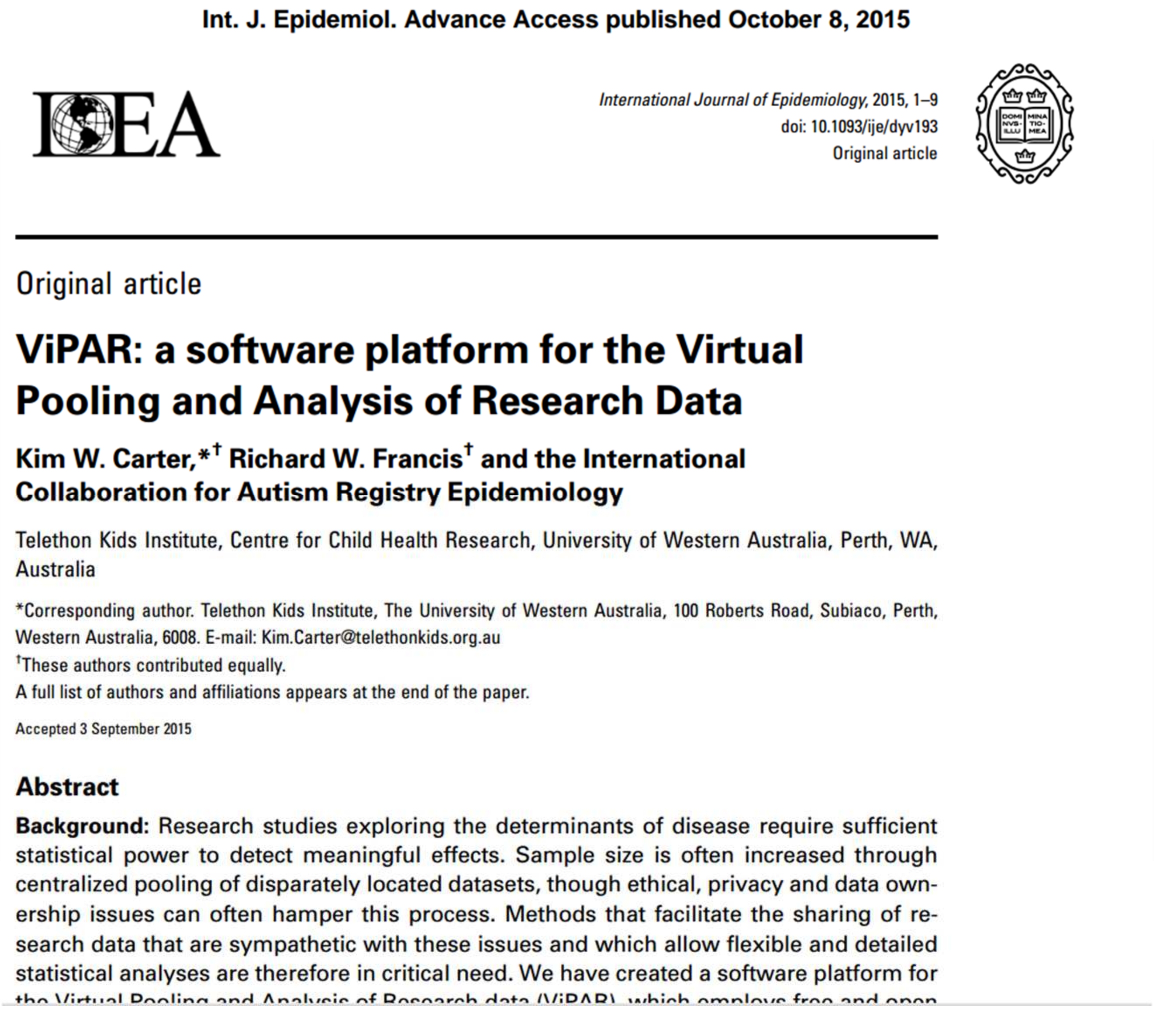
⊕ Author information

Abstract

Advancing paternal and maternal age have both been associated with risk for autism spectrum disorders (ASD). However, the shape of the association remains unclear, and results on the joint associations is lacking. This study tests if advancing paternal and maternal ages are independently associated with ASD risk and estimates the functional form of the associations. In a population-based cohort study from five countries (Denmark, Israel, Norway, Sweden and Western Australia) comprising 5 766 794 children born 1985-2004 and followed up to the end of 2004-2009, the relative risk (RR) of ASD was estimated by using logistic regression and splines. Our analyses included 30 902 cases of ASD. Advancing paternal and maternal age were each associated with increased RR of ASD after adjusting for confounding and the other parent's age (mothers 40-49 years vs 20-29 years, RR=1.15 (95% confidence interval (CI): 1.06-1.24), P-value<0.001; fathers⩾50 years vs 20-29 years, RR=1.66 (95% CI: 1.49-1.85), P-value<0.001). Younger maternal age was also associated with increased risk for ASD (mothers <20 years vs 20-29 years, RR=1.18 (95% CI: 1.08-1.29), P-value<0.001). There was a joint effect of maternal and paternal age with increasing risk of ASD for couples with increasing differences in parental ages. We did not find any support for a modifying effect by the sex of the offspring. In conclusion, as shown in multiple geographic regions, increases in ASD was not only limited to advancing paternal or maternal age alone but also to differences parental age including younger or older similarly aged parents as well as disparately aged parents.Molecular Psychiatry advance online publication, 9 June 2015; doi:10.1038/mp.2015.70.

Centrepiece of a successful NIH grant bid (2013-2017) $5.5M

For more information



Int. J. Epidemiol. Advance Access published October 8, 2015

IJEA

International Journal of Epidemiology, 2015, 1–9
doi: 10.1093/ije/dyv193
Original article

Original article

ViPAR: a software platform for the Virtual Pooling and Analysis of Research Data

Kim W. Carter,*[†] Richard W. Francis[†] and the International Collaboration for Autism Registry Epidemiology

Telethon Kids Institute, Centre for Child Health Research, University of Western Australia, Perth, WA, Australia

*Corresponding author. Telethon Kids Institute, The University of Western Australia, 100 Roberts Road, Subiaco, Perth, Western Australia, 6008. E-mail: Kim.Carter@telethonkids.org.au

[†]These authors contributed equally.

A full list of authors and affiliations appears at the end of the paper.

Accepted 3 September 2015

Abstract

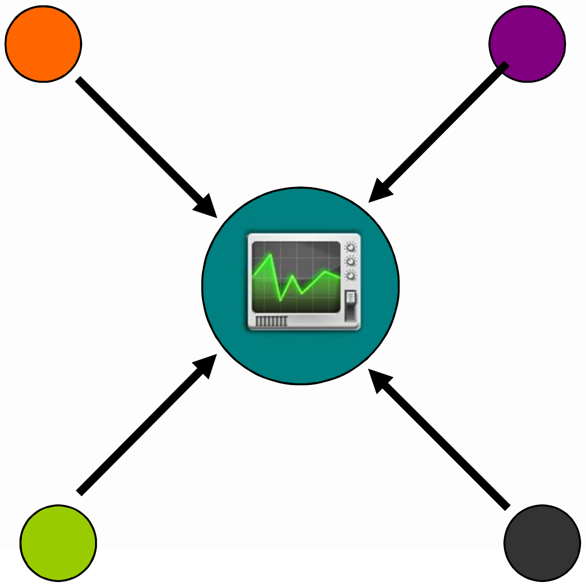
Background: Research studies exploring the determinants of disease require sufficient statistical power to detect meaningful effects. Sample size is often increased through centralized pooling of disparately located datasets, though ethical, privacy and data ownership issues can often hamper this process. Methods that facilitate the sharing of research data that are sympathetic with these issues and which allow flexible and detailed statistical analyses are therefore in critical need. We have created a software platform for the Virtual Pooling and Analysis of Research data (ViPAR), which employs free and open

Int J Epi 2015; doi: 10.1093/ije/dyv193

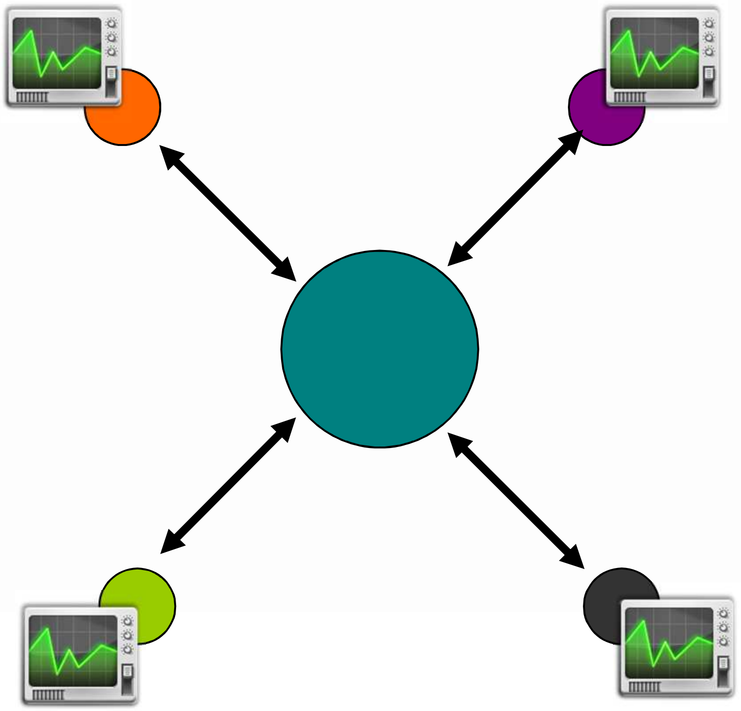
TELETHON KIDS INSTITUTE
Discover. Prevent. Cure.

ViPAR vs DataSHIELD

ViPAR



DataSHIELD



ViPAR vs DataSHIELD

- Data remains at study site
- Harmonisation is required before copies of data are sent ‘to the cloud’
- Generation of pooled dataset occurs ‘in the cloud’
- Results generated from a combined virtual dataset
- Pooled virtual dataset not saved/stored – only exists while analysis is running
- Straightforward analysis syntax in multiple languages

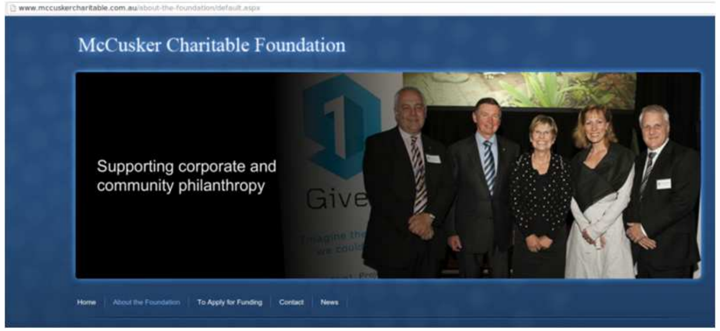
- Data remains at the study site
- Harmonisation is required before sending
- No data is sent anywhere
- No pooled dataset is generated
- Statistical methods sent to each study site, where methods are run locally
- Results are pooled back at analysis centre, to produce final result
- Requires specific analysis syntax and software (R) to make this happen

Future Directions

- Transparent integration of ViPAR with other complementary techniques eg DataSHIELD
- Web 2.0 interface
- Building capabilities to handle complex and large data types eg imaging data, whole genome seq
  - NoSQL, Object and other DBs
- Further security enhancements
  - on the fly privacy checks
  - using containers (Docker) for running analyses for greater separation

Acknowledgements

- Richard Francis, TKI


www.mccuskercharitable.com.au/about-the-foundation/default.aspx
Give
Home
About the Foundation
To Apply for Funding
Contact
News

McCusker Charitable Foundation

Supporting corporate and community philanthropy

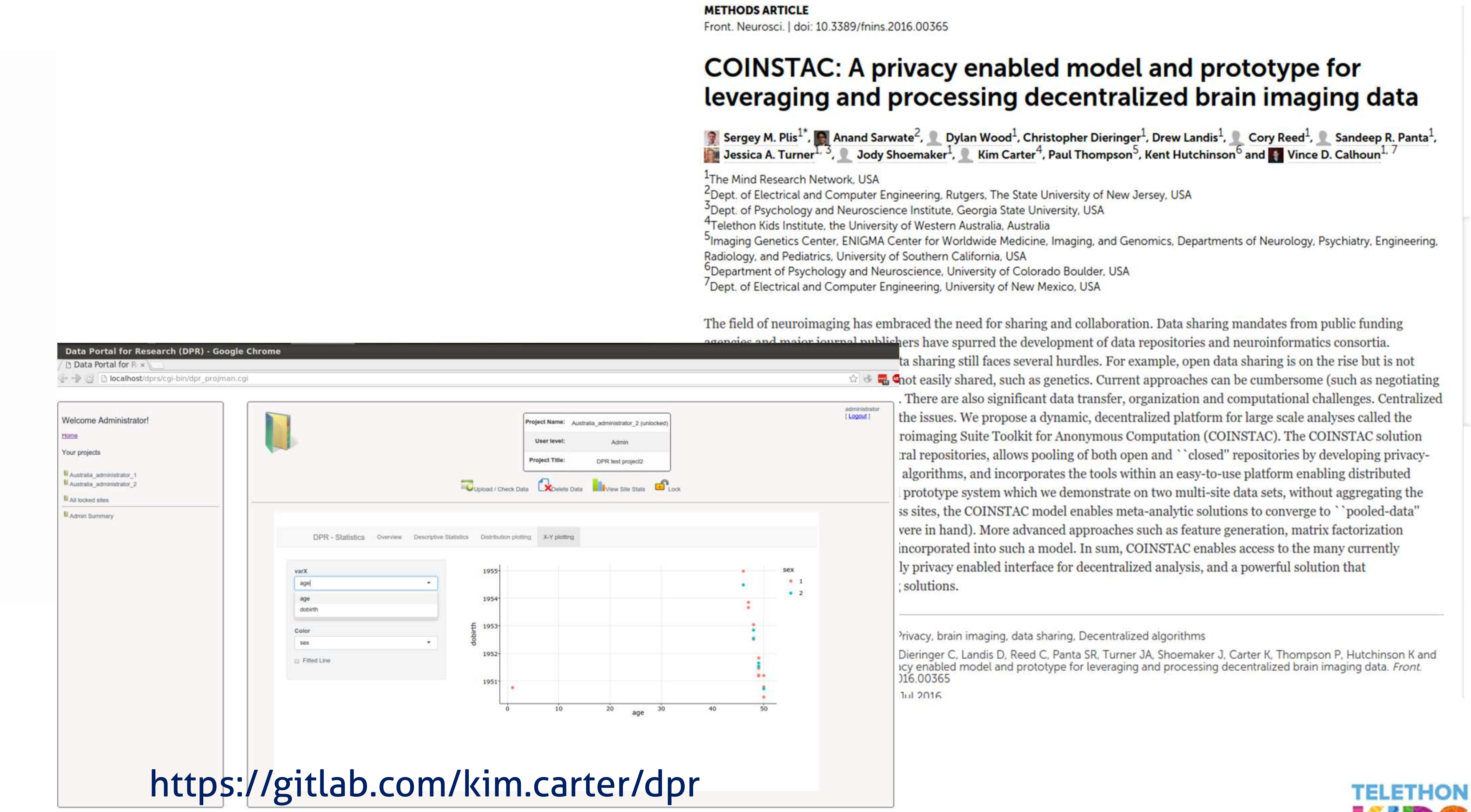
AUTISM SPEAKS™
It's time to listen.

NIH
National Institutes of Health

- Becca Wilson, Univ of Bristol
- Paul Burton, Univ. of Bristol

WUN
WORLDWIDE UNIVERSITIES NETWORK

A quick plug



METHODS ARTICLE
Front. Neurosci. | doi: 10.3389/fnins.2016.00365

COINSTAC: A privacy enabled model and prototype for leveraging and processing decentralized brain imaging data

Sergey M. Plis[1*], Anand Sarwate[2], Dylan Wood[1], Christopher Dieringer[1], Drew Landis[1], Cory Reed[1], Sandeep R. Panta[1], Jessica A. Turner[1, 3], Jody Shoemaker[1], Kim Carter[4], Paul Thompson[5], Kent Hutchinson[6] and Vince D. Calhoun[1, 7]

[1]The Mind Research Network, USA
[2]Dept. of Electrical and Computer Engineering, Rutgers, The State University of New Jersey, USA
[3]Dept. of Psychology and Neuroscience Institute, Georgia State University, USA
[4]Telethon Kids Institute, the University of Western Australia, Australia
[5]Imaging Genetics Center, ENIGMA Center for Worldwide Medicine, Imaging, and Genomics, Departments of Neurology, Psychiatry, Engineering, Radiology, and Pediatrics, University of Southern California, USA
[6]Department of Psychology and Neuroscience, University of Colorado Boulder, USA
[7]Dept. of Electrical and Computer Engineering, University of New Mexico, USA

The field of neuroimaging has embraced the need for sharing and collaboration. Data sharing mandates from public funding agencies and major journal publishers have spurred the development of data repositories and neuroinformatics consortia.
ta sharing still faces several hurdles. For example, open data sharing is on the rise but is not
not easily shared, such as genetics. Current approaches can be cumbersome (such as negotiating
. There are also significant data transfer, organization and computational challenges. Centralized
the issues. We propose a dynamic, decentralized platform for large scale analyses called the
roimaging Suite Toolkit for Anonymous Computation (COINSTAC). The COINSTAC solution
:ral repositories, allows pooling of both open and ``closed'' repositories by developing privacy-
algorithms, and incorporates the tools within an easy-to-use platform enabling distributed
prototype system which we demonstrate on two multi-site data sets, without aggregating the
ss sites, the COINSTAC model enables meta-analytic solutions to converge to ``pooled-data''
vere in hand). More advanced approaches such as feature generation, matrix factorization
incorporated into such a model. In sum, COINSTAC enables access to the many currently
ly privacy enabled interface for decentralized analysis, and a powerful solution that
; solutions.

?rivacy, brain imaging, data sharing, Decentralized algorithms

Dieringer C, Landis D, Reed C, Panta SR, Turner JA, Shoemaker J, Carter K, Thompson P, Hutchinson K and
ıcy enabled model and prototype for leveraging and processing decentralized brain imaging data. *Front.*
)16.00365

Jul 2016

Data Portal for Research (DPR) - Google Chrome

localhost/dprs/cgi-bin/dpr_projman.cgi

Welcome Administrator!

Home

Your projects

- Australia_administrator_1
- Australia_administrator_2
- All locked sites
- Admin Summary

Project Name:	Australia_administrator_2 (unlocked)
User level:	Admin
Project Title:	DPR test project2

varX: age

Color: sex

□ Fitted Line

https://gitlab.com/kim.carter/dpr

Questions & Demo time?

ÁLA'IH, DO'NEH'LINI, ÁLA'IH, DO'NEH'LINI, ÁLA'IH, DO'NEH'LINI, ÁLA'IH, DO'NEH'LINI, DO'NEH'LINI, ÁLA'IH, DO'NEH'LINI, ÁLAIH, DO'NEH'LINI, ÁLA'IH, DO'NEH'LINI, DO'NEH'LINI, DO'NEH'LINI...
FOR ADDED SECURITY, AFTER WE ENCRYPT THE DATA STREAM, WE SEND IT THROUGH OUR NAVAJO CODE TALKER.
...IS HE JUST USING NAVAJO WORDS FOR "ZERO" AND "ONE"?
WHOA, HEY, KEEP YOUR VOICE DOWN!

http://xkcd.com/257/