Moving Biomedical Big Data Sharing Forward

An adoption of the RDA Data Citation of Evolving Data Recommendation to Electronic Health Records

Leslie McIntosh
Cynthia Hudson Vitale

Daniel Vianello

Snehil Gupta
Connie Zabarovskaya
Brian Romine

January 2017

Background

Director, Center for Biomedical Informatics

Leslie McIntosh, PHD, MPH

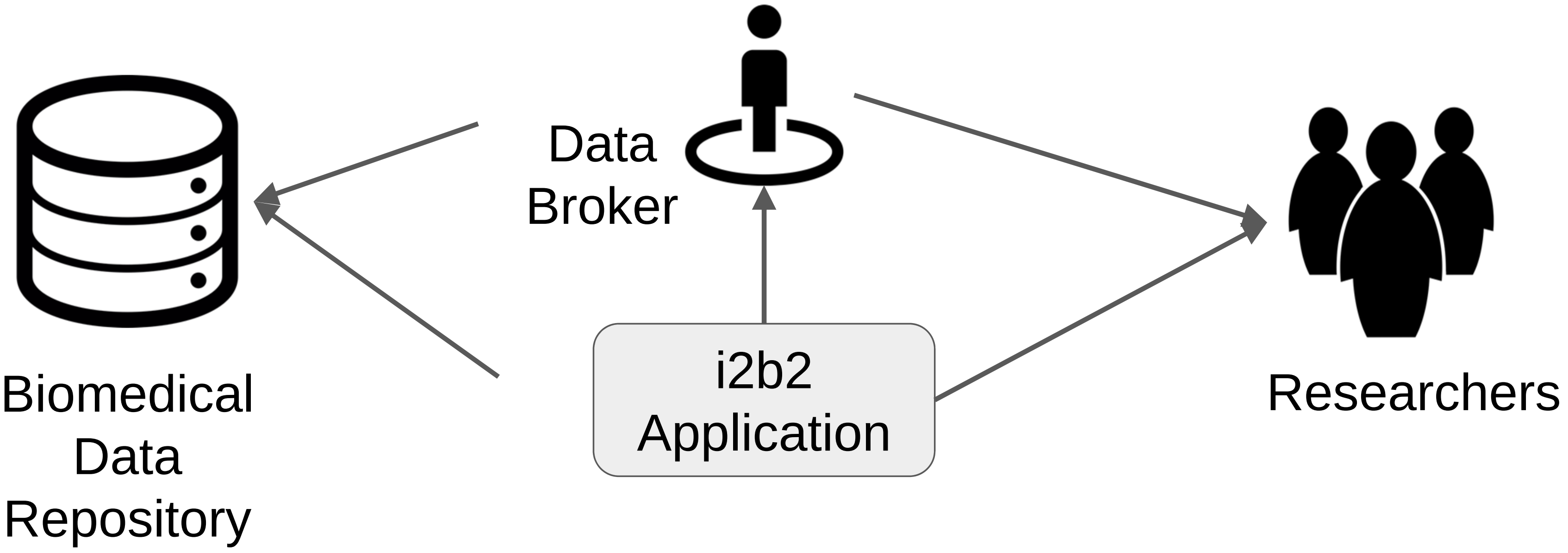
Data Services Coordinator

Cynthia Hudson Vitale, MA

Washington University in St.Louis

BDaaS

Biomedical Data as a Service



Move some of the responsibility of reproducibility

Biomedical Researcher → Biomedical Pipeline

RDA/MacArthur Grant

Biomedical Adoption Project Goals

- Implement RDA Data Citation WG recommendation to local Washington U i2b2
- Engage other i2b2 community adoptees
- Contribute source code back to i2b2 community

RDA Data Citation WG Recommendations

- R1: Data Versioning
- R2: Data Timestamping
- R3, R9: Query Store
- R7: Query Timestamping
- R8: Query PID
- R10: Query Citation

Internal Implementation Requirements

- Scalable
- Available for PostgreSQL
- Actively supported
- Easy to maintain
- Easy for data brokers to use

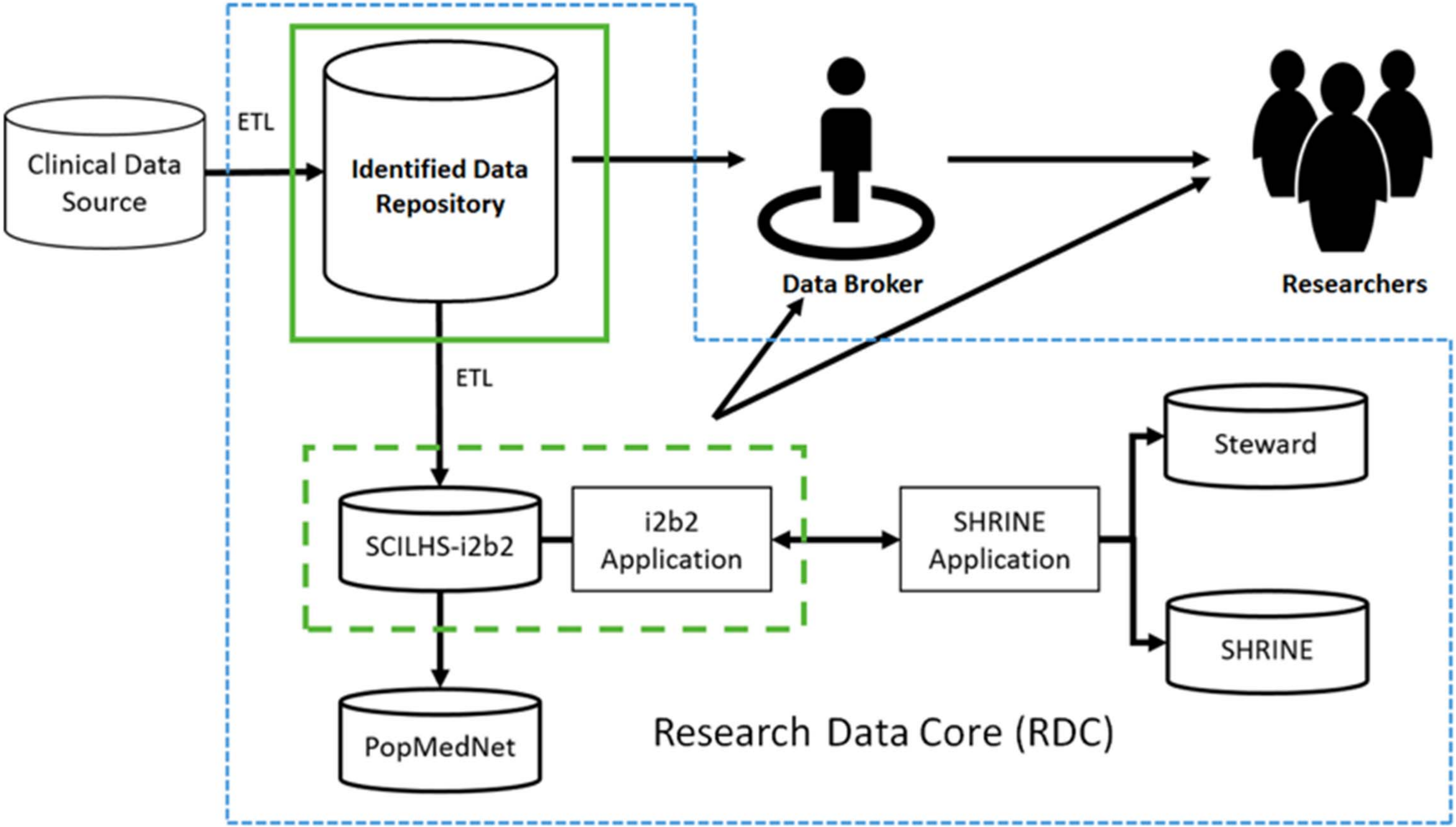

Clinical Data Source
ETL
Identified Data Repository
Data Broker
Researchers
ETL
SCILHS-i2b2
i2b2 Application
SHRINE Application
Steward
SHRINE
PopMedNet
Research Data Core (RDC)

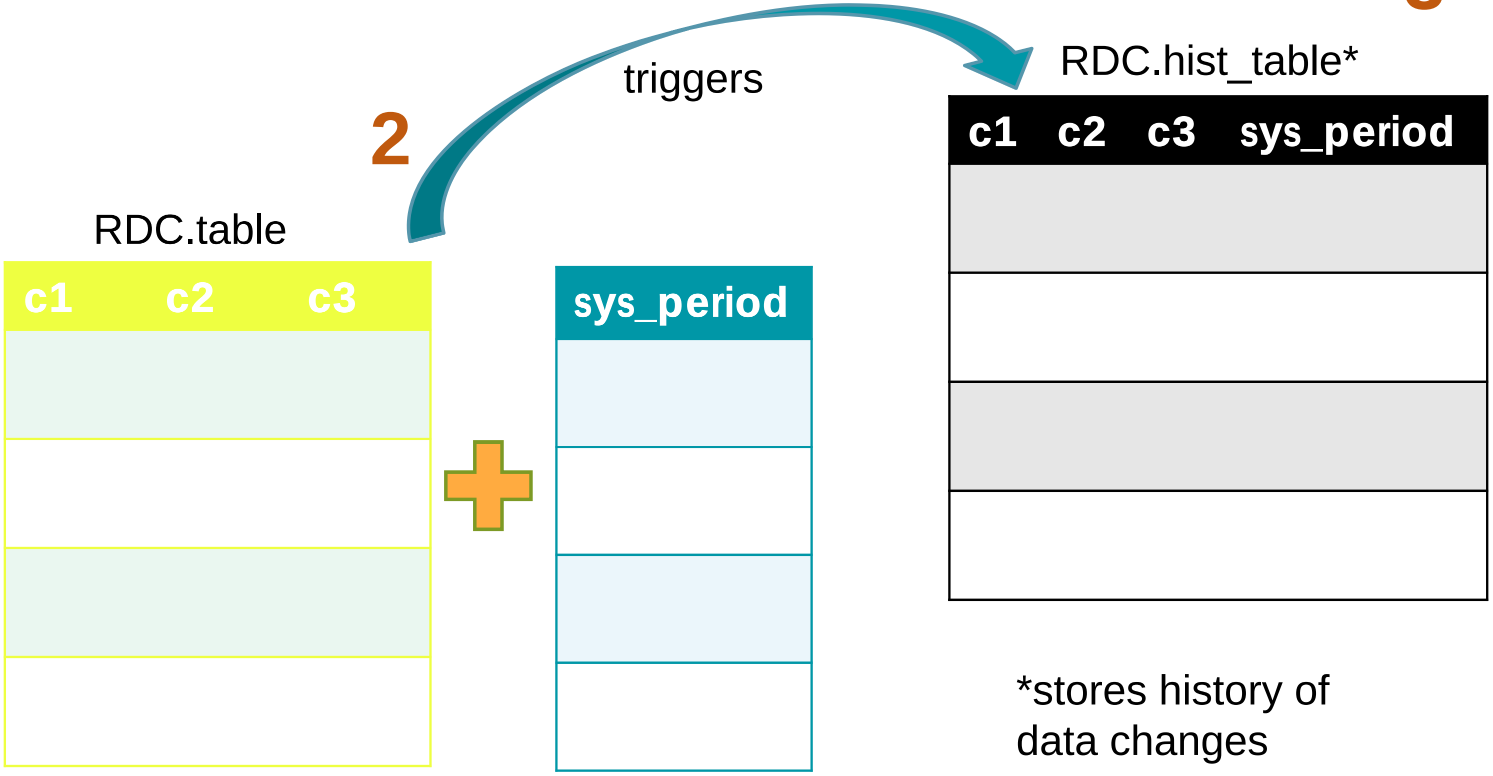
Table 2: Gap Analysis Summary

Database	**Data Versioning (R1)**	**Data Timestamp (R2)**	**Query Store (R3/R9)**	**Query Timestamp (R7)**	**Query PID (R8)**	**Citation Text (R10)**
Identified Data Repository	Yes (default)	Yes (default)	No	No	No	No
i2b2 (Local)	No	No	Yes (i2b2 default)	Yes (i2b2 default)	Yes (i2b2 default)	No
i2b2 (SCILHS)	No	No	Yes (i2b2 default)	Yes (i2b2 default)	Yes (i2b2 default)	No
PopMedNet	No	No	No	No	No	No

R1 and R2 Implementation

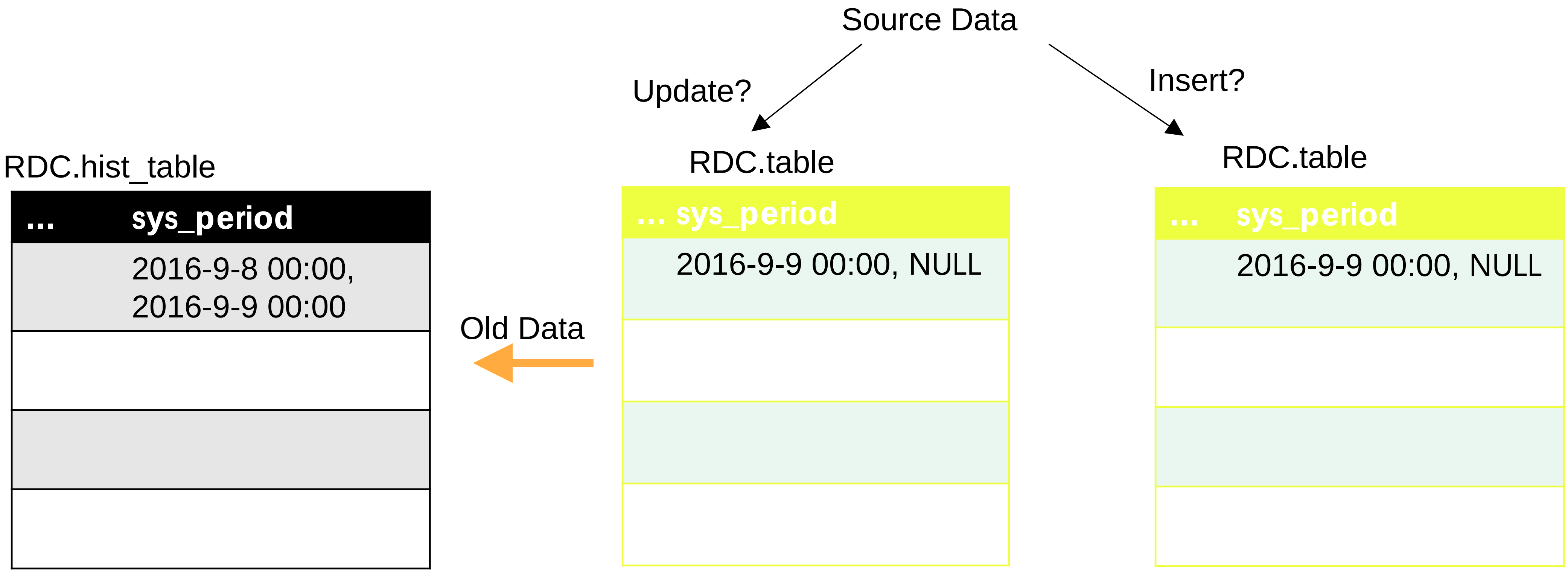
1

PostgreSQL Extension
“temporal_tables”



*stores history of
data changes

ETL Incrementals



Source Data

Update?

Insert?

Old Data

RDC.hist_table

...	sys_period
	2016-9-8 00:00, 2016-9-9 00:00

RDC.table

...	sys_period
	2016-9-9 00:00, NULL

RDC.table

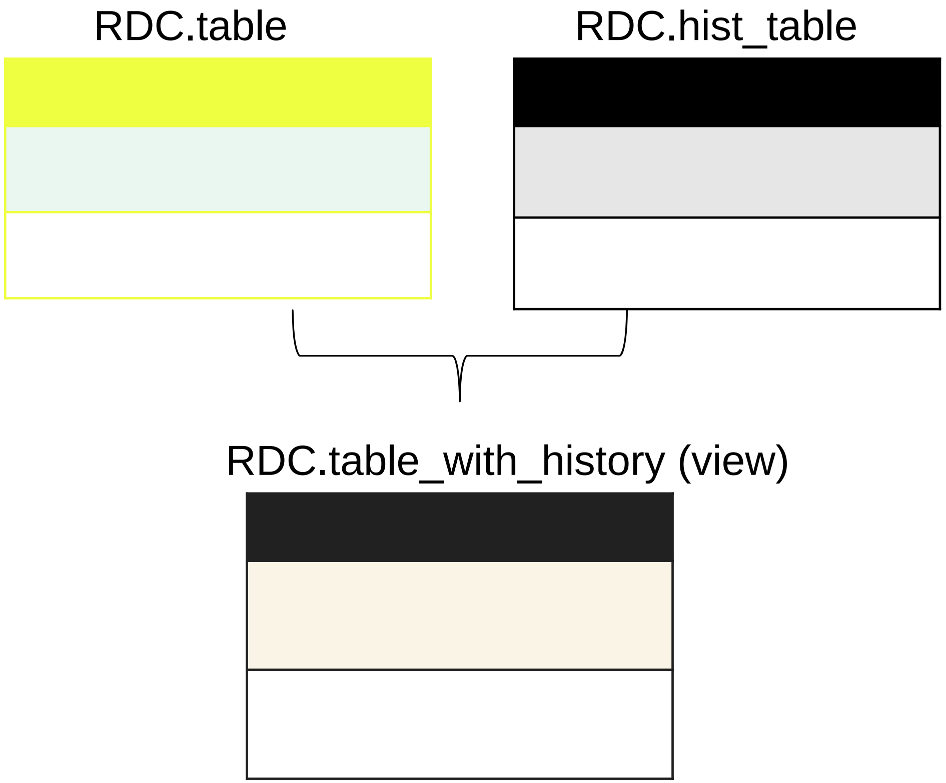
...	sys_period
	2016-9-9 00:00, NULL

R3, R7, R8, R9, and R10 Implementation

1

PostgreSQL Extension
“temporal_tables”

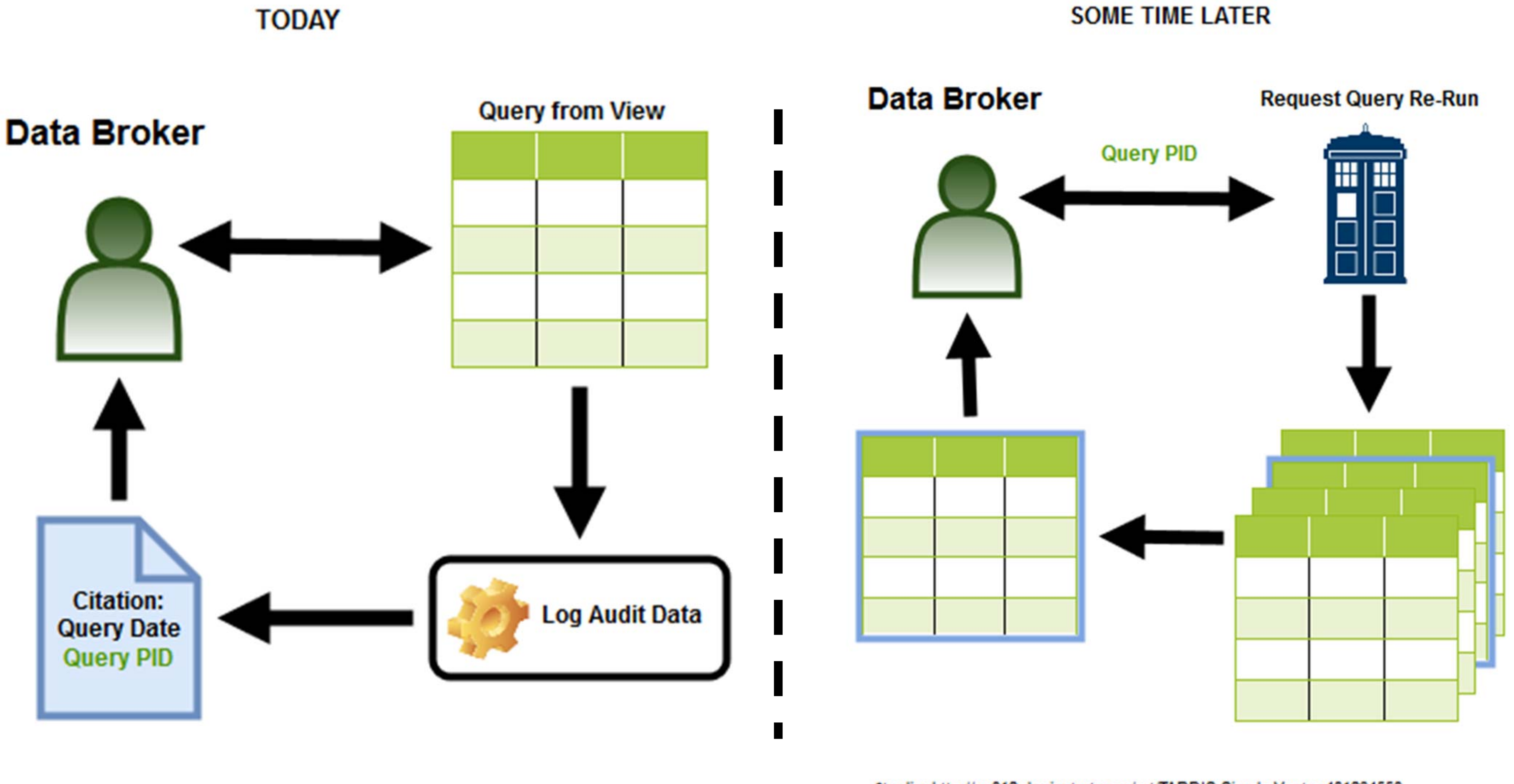
2



3

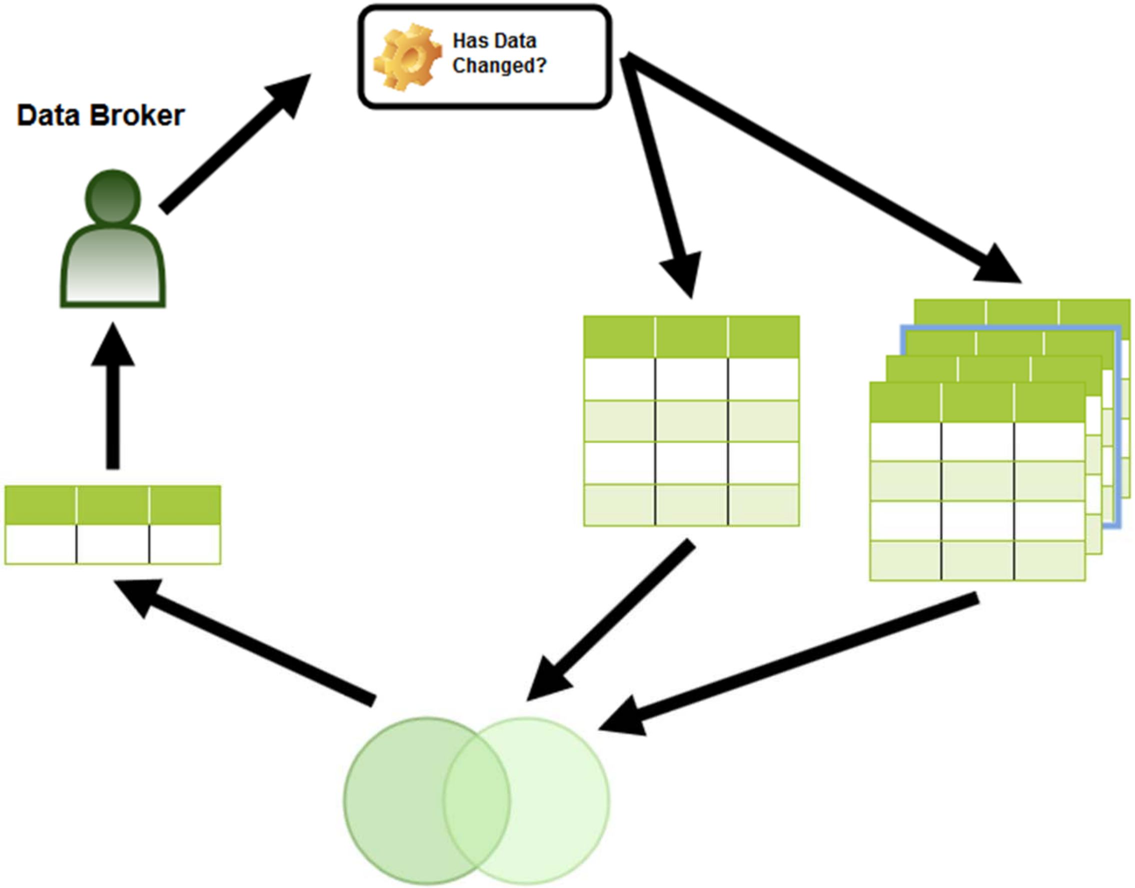
- functions
- triggers
- query audit tables

Data Reproducibility Workflow



*tardis: http://pc012.deviantart.com/art/TARDIS-Simple-Vector-481264558

Bonus Feature: Determine if Change Occurred



Future Developments

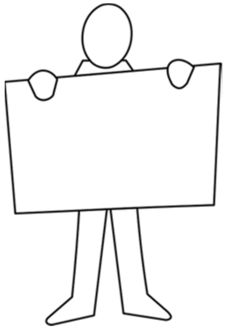
- Develop a process for sharing Query PID with researchers in an automated way
- Resolve Query PIDs to a landing page with Query metadata
- Implement research reproducibility requirements in other systems as possible

Outcomes and Support

Obtained Outcomes

- Implemented WG recommendations
- Engaged with other i2b2 adoptees
  (Harvard, Nationwide Children’s Hospital)

Dissemination



- Poster presentation (Harvard U, July 2016)
- Scientific manuscript based on our proof of concept to AMIA TBI/CRI 2017 conference
- Sharing the code with the community

Return on Investment (ROI) - Estimated

- 20 hours to complete 1 study
- $150/hr (unsubsidized)
- $3000 per study
- 115 research studies per year
- **14 replication studies**

Funding Support

MacArthur Foundation 2016 Adoption Seeds program Foundation through a sub-contract with Research Data Alliance

Washington University Institute of Clinical and Translational Sciences
NIH CTSA Grant Number UL1TR000448 and UL1TR000448-09S1

Siteman Cancer Center at Washington University
NIH/NCI Grant P30 CA091842-14

Center for Biomedical Informatics @WUSTL
Teams for Reproducible Research

NIH-NLM Supplement

Leslie McIntosh
Cynthia Hudson-Vitale
Anthony Juehne
Rosalia Alcoser
Xiaoyan ‘Sean’ Liu
Brad Evanoff

RDA Collaborators

Andreas Rauber
Stefan Pröll

Research Data Alliance

Leslie McIntosh
Cynthia Hudson-Vitale
Anthony Juehne
Snehil Gupta
Connie Zabarovskaya
Brian Romine
Dan Vianello

Washington University in St.Louis

WashU CBMI Research Reproducibility Resources

Repository

https://github.com/CBMIWU/Research_Reproducibility

Slides

http://bit.ly/2cnWorU

Bibliography

https://www.zotero.org/groups/biomedical_informatics_resrepro