



13 March 2014

Decomposing Results Without Burying the Body of Evidence: A Modus Operandi
for Developing Metadata and Digital Preservation Requirements

Lorraine L. Richards Bornn, PI

Decomposing Results Without Burying the Body of Evidence:

A Modus Operandi for Developing Metadata and
Digital Preservation Requirements



DREXEL UNIVERSITY

College of

Computing & Informatics

Dr. Lorraine L. Richards, Assistant Professor

Adam Townes, Doctoral Candidate

Dr. William C. Regli, Professor

YuanYuan Feng, Doctoral Student

Acknowledgments

We want to thank the FAA's William J. Hugh's Technical Center and its personnel for allowing us access to their labs and employees in the pursuance of this project.



Agenda

- Introduction – Scientific Data Reuse, Sharing and Preservation
- Relationship to current research in digital/data curation
- Description of project and its goals
- Project venue and environment
- Selection of laboratories for the project
- Challenges of collaboratively developing the metadata



Introduction

- Greater collaboration among scientific researchers
 - The “fourth paradigm” of science – data intensive (Grey 2007)
- Big data sets
- Increasing desires to repurpose data and to ensure reproducibility of results
- This has become a big issue for the Federal Government, with Obama’s Open Data Policy
 - Has resulted in a memoranda and a policy document requiring federal agencies spending more than \$100 million on scientific research to create systems that will support the sharing of data and results outside of their particular agencies



Relationship to current research

- A key requirement for scientists is *reproducibility* (Faniel & Jacobsen 2010)
 - *Reproducibility* relies on ensuring that the data is
 - Discoverable
 - Trustworthy (i.e., it is accurate and has not been modified in unknown ways since the last time the experiment was performed)
- We want to describe not only the research objects that contain the aggregated information, but also use metadata to describe the individual components that comprise the objects! (Bechhofer et al.
 - Data, results of analysis, goals of the experiment or simulation, how data was created and modified throughout the scientific process, experimental parameters, intermediate results, logs, final results, and even problems encountered when creating the data for use
- Tracking provenance is crucial! (Muniswamy-Reddy et al. 2009).



Our Project – FAA's William J. Hughes Technical Center



Photos from: Federal Aviation Administration. (2004). *Welcome to William J. Hughes Technical Center.*

http://www.faa.gov/about/office_org/headquarters_offices/ang/offices/tc/.

Work at the FAA Tech Center

- Aviation scientific research facility
- Testing and Development of new and existing
 - Equipment
 - Systems
 - Procedures
 - Materials
- Using simulations and experiments to test and improve safety in the NAS



The Continuing Mission of the FAA:

**To provide the safest, most efficient
aerospace system in the world.**

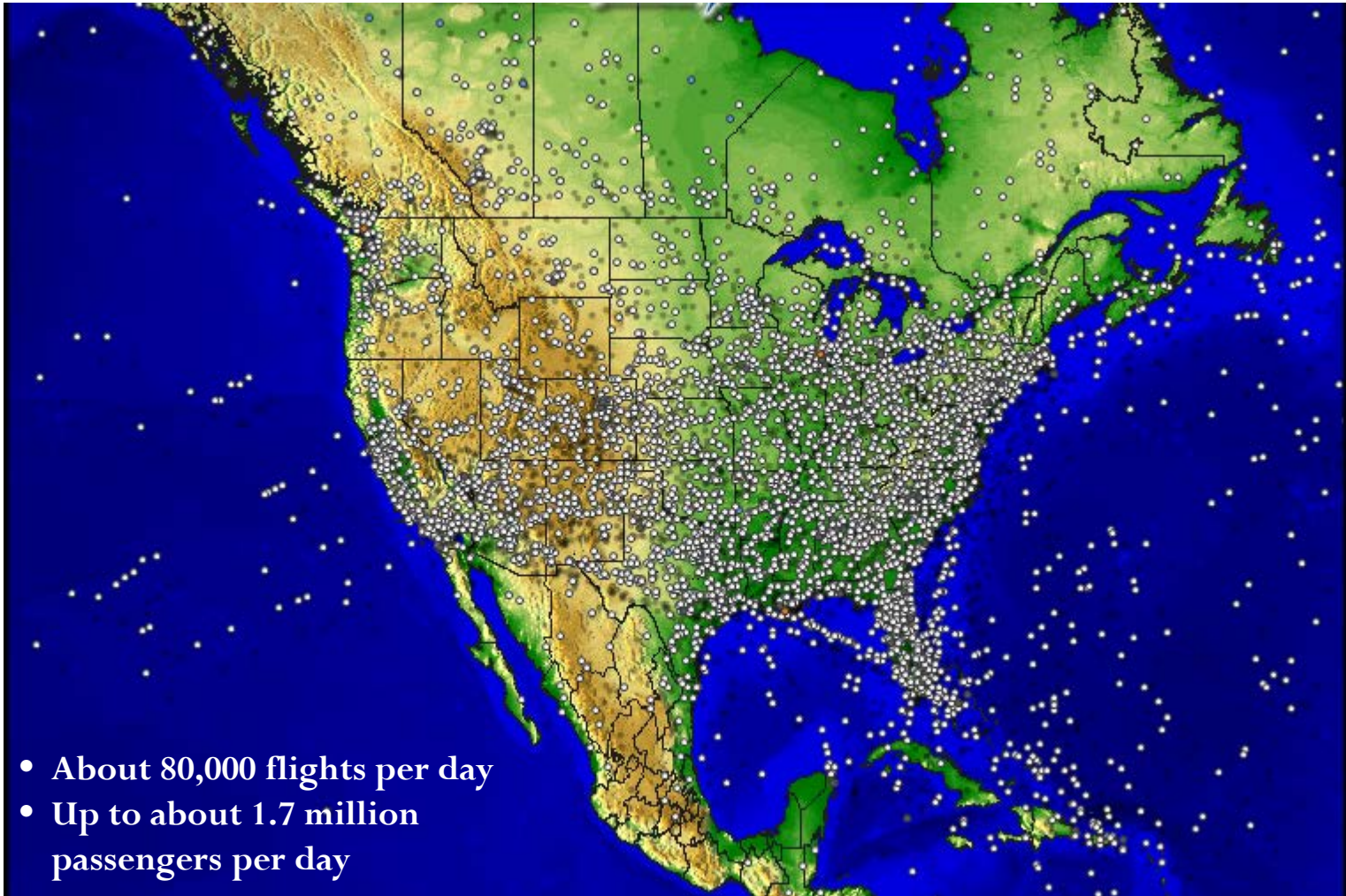


The FAA Achieves Air Safety Success

- The U.S. National Air Space (NAS) *is* among the safest in the world, according to the International Civil Aviation Organization (ICAO), a specialized agency of the United Nations that sets air traffic standards and protocols worldwide (ICAO 2012, 10).

UN Region	Accidents	Fatalities	% Accidents	% Fatal Accidents	% Fatalities
Africa	5	167	5%	22%	45%
Asia	23	161	23%	33%	43%
Europe	30	42	30%	33%	11%
Latin America & the Caribbean	12	2	12%	12%	1%
Northern America	29	0	30%	0%	0%
Oceania	0	0	0%	0%	0%
World	99	372			

U.S. Air Traffic - Facts



Tech Center Data Environment “As-Is”

- Numerous scientific labs spanning a variety of disciplines
- “Big Data” environment
 - We received a dataset for *one experiment* from *one lab* that was over 2.5 Terabytes in size
- Very little previous sharing of data
- Management of research results sets “siloed”
 - Reuse is rare; when it occurs, a researcher needs to personally contact the original Project Investigator (PI)
- They are not yet “curation-centric” and have little knowledge of preservation requirements



Our goals

- Creation and enhancement of current data sets and sources
- Mutually produced project plan for enabling scenario creation with reusable data
- Understanding of current and required metadata
- Development of a metadata taxonomy
- Design of a prototype that exhibits the capabilities of auto-generating metadata tags
- Specification for rules and policies for the data sets and for access controls



To-Be: A “best practice” approach for a big data scientific laboratory

- Getting there requires:
 - Addressing automated search and discovery
 - Determining which data is most valuable for reuse by a diverse set of scientists
 - Incentivizing sharing and reuse
 - Ensuring trust in the data
 - Assessing and meeting preservation requirements

Cultural change and introduction of a “curation-centric” mindset is typically the most difficult part of a project like this. This project is no exception.



Steps

- Understanding business (i.e., scientific research) context
- Understanding nature of the data
- Developing new scenarios for simulations and experiments
- Working collaboratively to assess relevant metadata
- Developing prototype federated system
- *Conducting continuous, clarificatory communication with the project's partner institution*
 - *Explaining the benefits of engaging in this activity (and then re-explaining throughout the project)*
 - *Taking an iterative approach to the explanations*



Challenges creating big data metadata

- Large volume precludes individual data element-level “search and discover” tactics
 - Auto-generation of metadata and automated, rule-based functionality become much more important
- After determining appropriate subset for analysis, you “discover” which data elements are those that are relevant
 - Highly collaborative undertaking
 - Relies upon having a good process for adding data elements, since new experiments/scenarios will require additional data to be added to the repository on an incremental basis
 - Relies on the subject-matter experts
 - Tracking versions of data since it is undesirable to modify the original dataset used for a new experiment
- Maintaining provenance of data, analyses and results from scientific research



The Participants: Three Labs

- Explain the three labs – merge inform from “Selecting Labs” directly into this concrete description of the three labs
 - Human Factors
 - Target Generation Facility
 - The Wildlife Hazard group within the Airport Safety R&D Division



What is a Simulation?

- Traffic Flow Management
 - Management of the flow of air traffic in the National Airspace System is based on capacity and demand.
- Time Based Flow Management (TBFM)
 - TBFM is the technology and method used for adjusting capacity/demand imbalances at select airports, departure fixes, arrival fixes and en route points across the NAS.
 - Picture air traffic as a very complex, four dimensional zipper



















Simulation Example

- Time Based Flow Management (TBFM)



Simulation Example

Traffic Flow: Airport Approach				
				
				
				
				
				

Traffic Flow: Airport Approach Overhead		
		
		
		
		

Potential Users of the Research Data

- FAA
 - Later reuse by originating lab
 - Other FAA Labs
- Other Federal Agencies
- Public
- Academia
- Private Industry

FAA personnel assume that initially the primary users will be fellow scientists within the various FAA laboratories

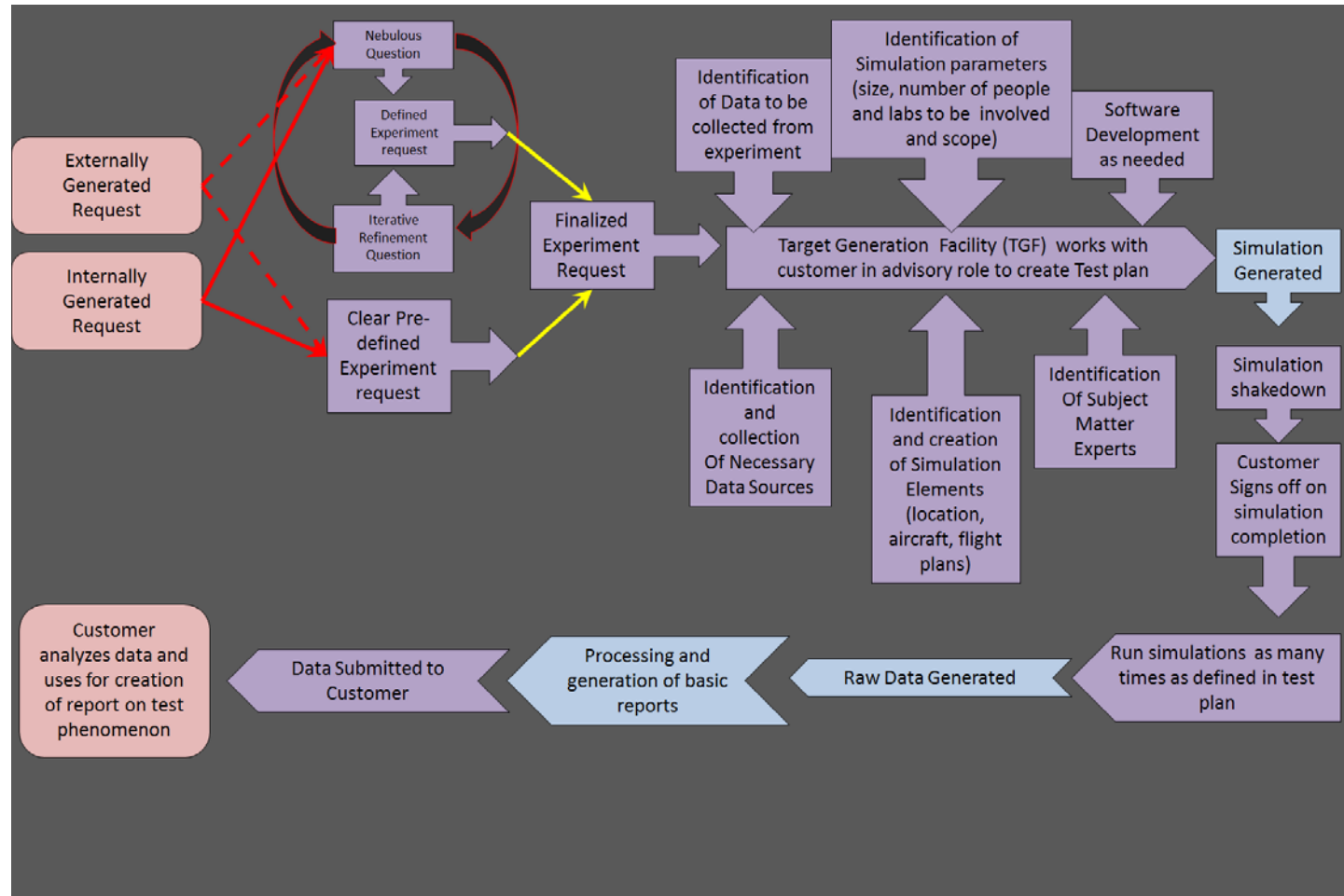


Understanding Workflow

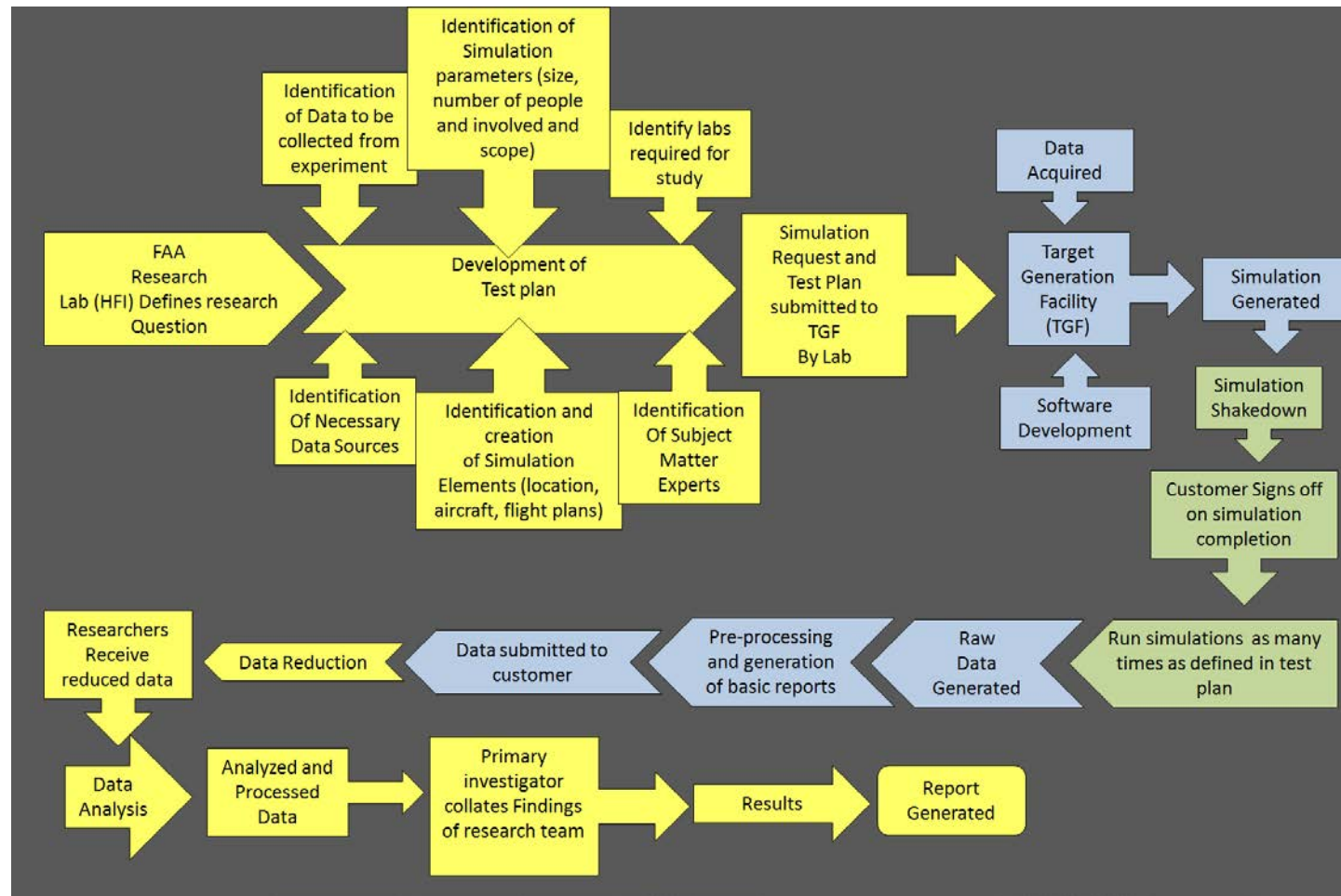
- Gradual Process of understanding
 - Visit site(s)
 - Shadow participants
 - Ask clarifying questions where feasible
 - Develop more detailed questions for interview protocol
 - Interview participants
 - Use semi-structured process
 - Asking clarifying questions as time allows
 - Draft diagram of process
 - Present diagram draft to same participants
 - Solicit feedback
 - Revise diagram according participant feedback



Workflow Model: Target Generation Facility



Workflow Model: Human Factors Lab



Developing the Metadata

- This is a collaborative “search and discover” process
- Marrying the scientific processes to the data and metadata
- Research (Faniel & Jacobsen 2010) has shown that for scientists to feel comfortable reusing data from other scientists they consider 3 main factors:
 - Relevance (do the existent data map to the potential research parameters? Do they use the same parameters?)
 - Understandability (is there enough documentation to ensure that the scientists know the precise way the data are defined and created and collected)
 - Trustworthiness (understanding how the data is produced increases trust, as does understanding how the previous scientists dealt with data-production problems.



Our Vision: Federated System

- Ingest data that is distributed across different groups and storage locations, including remote locations
 - Use of a Logical Name Space (a set of names used to describe entities in a consistent manner to the local user, regardless of source names)
- A metadata catalog allows local users to access local and remote data, through a rule-based process
 - Permanent database system that contains metadata mappings
- Allow the workflow to be executed at the site of the data
- Use rule-based assessment to insure that ingested data meets the metadata requirements
- Descriptive metadata at both aggregated object and property level allow discovery
- Preservation metadata includes
 - Provenance (authenticity) information
 - Representation information containing structure and semantics
 - Administrative information
- Retention, disposition, replication, access controls, checksums, etc.)



Our current steps

- Assessing the documentation and processes needed to enable the scientists to feel comfortable sharing and reusing data from other labs (and their own)
- Assessing the metadata and processes necessary for preservation over the longer term
- Creation of ontology and generation of OWL (and discovery/determination of required extensions)
- Building a prototype system, using iRODs technology, with basic Dublin Core.



Questions?



External References

- Bechhofer, S. et al. (2010). Why Linked Data is Not Enough for Scientists. Sixth IEEE e-Science Conference. Brisband, Australia.
<http://eprints.ecs.soton.ac.uk/21587/>.
- Faniel, Ixchel M. and Trond E. Jacobsen. (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. Computer Supported Cooperative Work (3-3/4): 355-375.
<http://link.springer.com/article/10.1007%2Fs10606-010-9117-8>.
- Hey, Tony et al. (2009). Jim Grey on e-Science: A Transformed Scientific Method. In The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA: Microsoft Research.
- International Civil Aviation Organization (ICAO). (2013). 2013 Safety Report. (Montreal: ICAO). http://www.icao.int/safety/Documents/ICAO_2013-Safety-Report_FINAL.pdf.



External References

- Muniswamy-Reddy et al. (2009). Making a Cloud Provenance-Aware. *Proceedings of TAPP'09, the First workshop on Theory and Practice of Provenance*. Berkeley, CA: USENIX Association.
<http://dl.acm.org/citation.cfm?id=1525944>. Rajasekar, Arcot et al. (2010). *iRODS Primer: integrated Rule-Oriented Data System*. na: Morgan & Claypool.
<http://www.morganclaypool.com/doi/pdfplus/10.2200/S00233ED1V01Y200912ICR012>.
- World Bank. (2014). *Air Transport, Passengers Carried*.
<http://data.worldbank.org/indicator/IS.AIR.PSGR/countries/1W?display=default>.
- World Bank. (2014). *Air Transport, Registered Carrier Departures Worldwide*.
<http://databank.worldbank.org/data/views/reports/tableview.aspx>.



Presenter Contact Information

Lorraine L. Richards, Assistant Professor, llr43@drexel.edu

Adam Townes, amt74@drexel.edu

