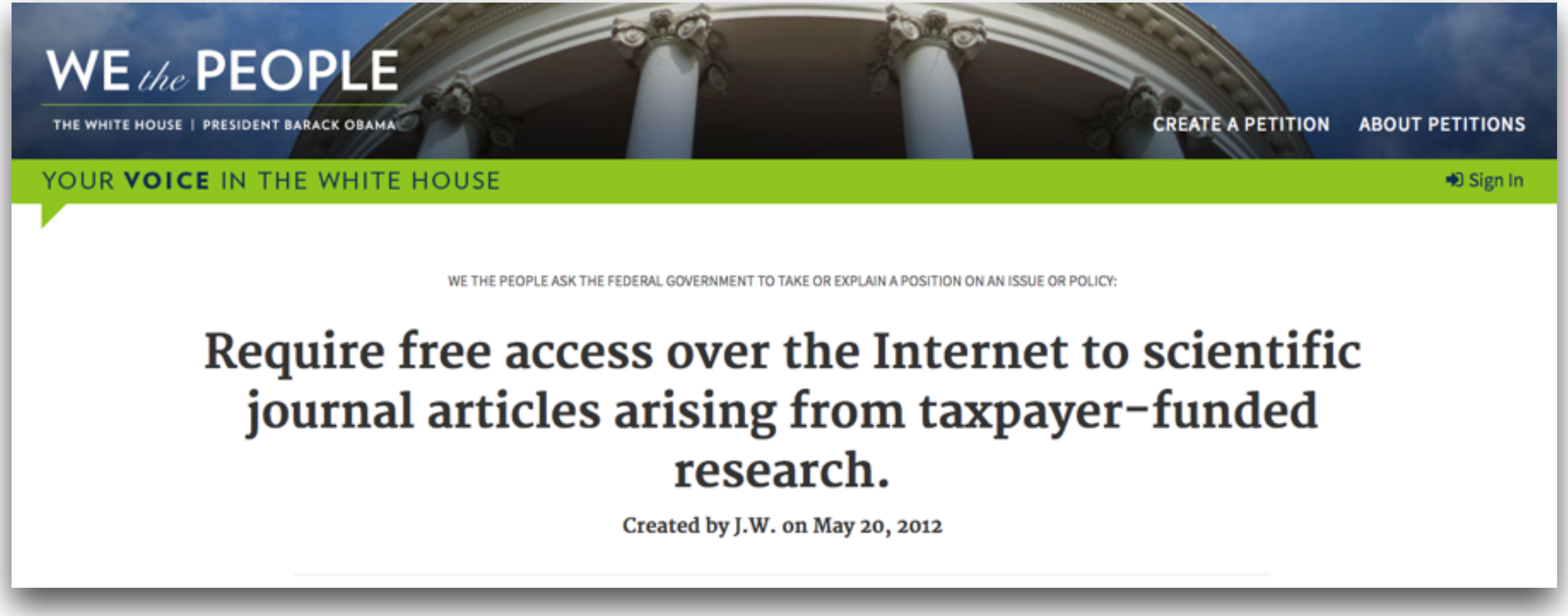
Costs of Curation

INFX 551
Data Curation

- Next week: Summary Lecture + Presentations
- Project Updates
- Lecture & Discussion
- Activity (s)
- Gordon Kennedy - WSDoT

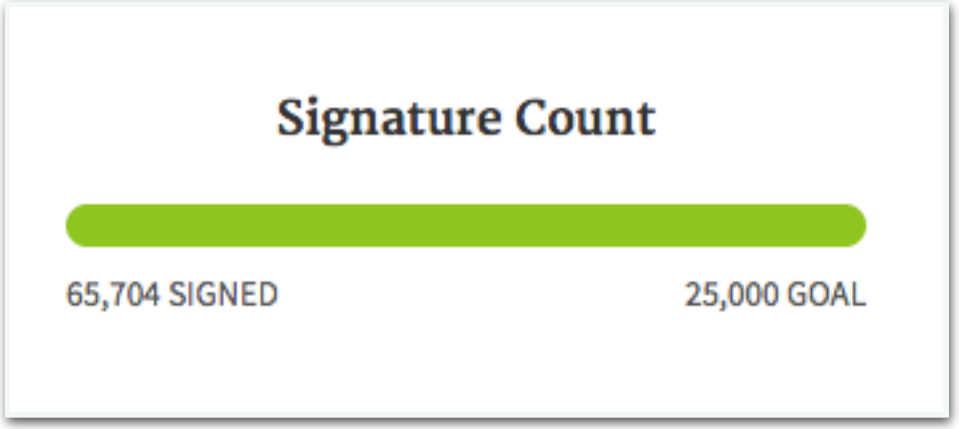
Policy background for open data in the USA…


WE the PEOPLE
THE WHITE HOUSE | PRESIDENT BARACK OBAMA
CREATE A PETITION
ABOUT PETITIONS
YOUR VOICE IN THE WHITE HOUSE
Sign In

WE THE PEOPLE ASK THE FEDERAL GOVERNMENT TO TAKE OR EXPLAIN A POSITION ON AN ISSUE OR POLICY:

Require free access over the Internet to scientific journal articles arising from taxpayer-funded research.

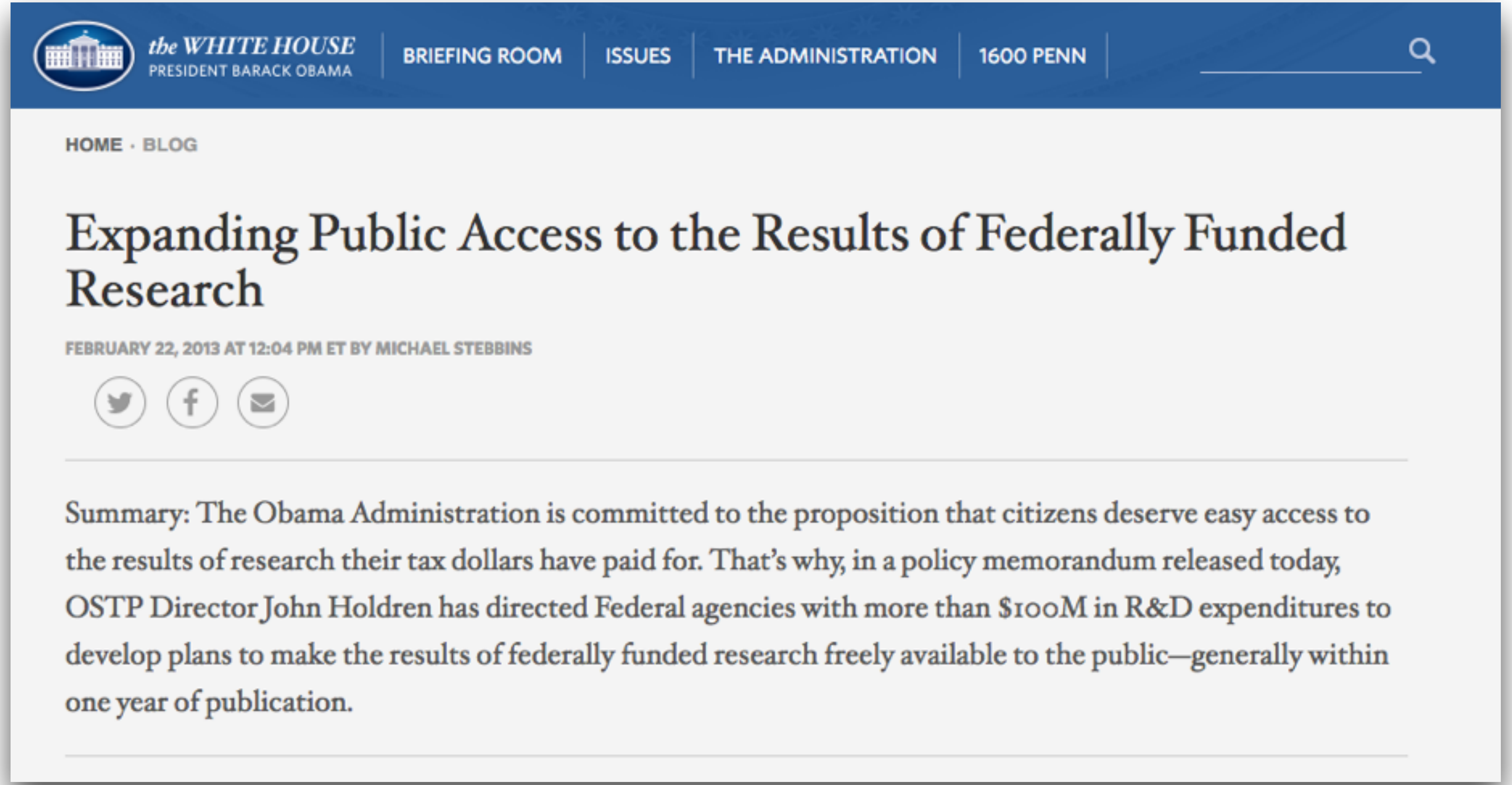
Created by J.W. on May 20, 2012



Signature Count

65,704 SIGNED

25,000 GOAL

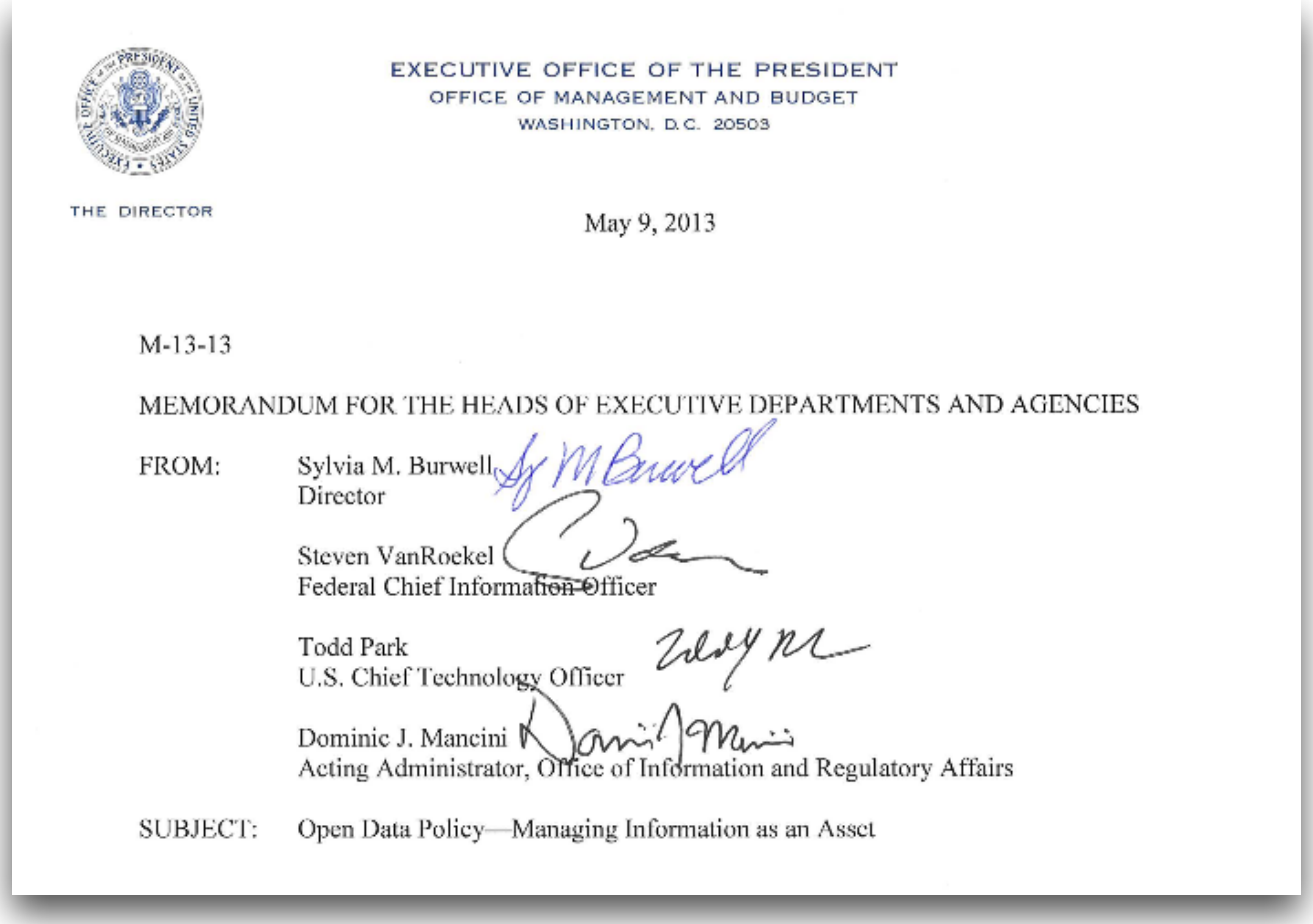


Expanding Public Access to the Results of Federally Funded Research

FEBRUARY 22, 2013 AT 12:04 PM ET BY MICHAEL STEBBINS

Summary: The Obama Administration is committed to the proposition that citizens deserve easy access to the results of research their tax dollars have paid for. That's why, in a policy memorandum released today, OSTP Director John Holdren has directed Federal agencies with more than $100M in R&D expenditures to develop plans to make the results of federally funded research freely available to the public—generally within one year of publication.

Read the memorandum here: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf



EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503

THE DIRECTOR

May 9, 2013

M-13-13

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Sylvia M. Burwell
Director

Steven VanRoekel
Federal Chief Information Officer

Todd Park
U.S. Chief Technology Officer

Dominic J. Mancini
Acting Administrator, Office of Information and Regulatory Affairs

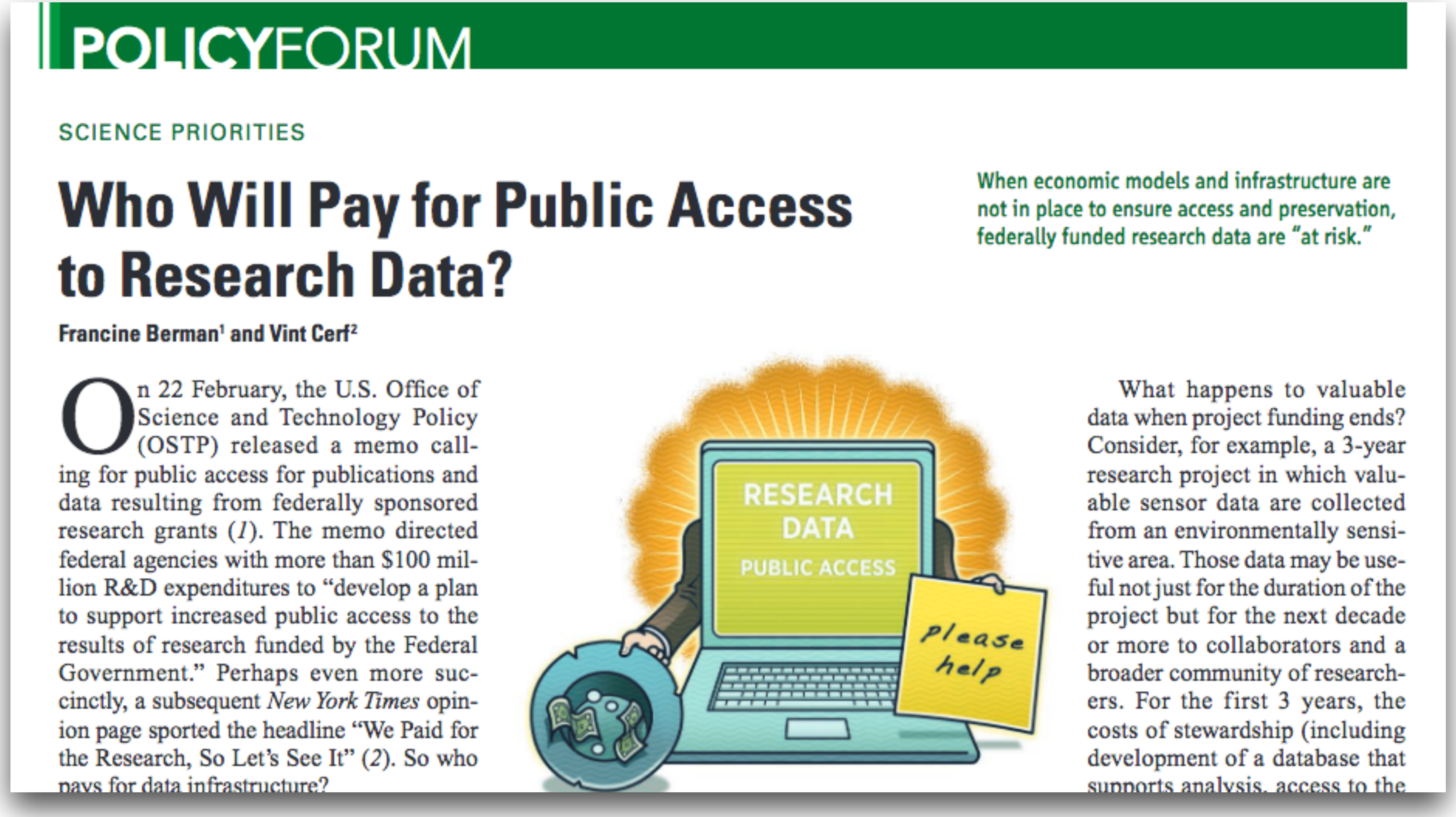
SUBJECT: Open Data Policy—Managing Information as an Asset

"…this Memorandum requires agencies to collect or create information in a way that supports downstream information processing and dissemination activities."

Impact is twofold:

- Unfunded set of mandates to federal agencies - both memoranda offer choices and broad time frame for compliance (and don’t spell out sanctions).

- Spurs innovation for compliance - the government often says it doesn’t want to pick winners or losers - but it does want the game to be played. This is how the open data game started.



POLICYFORUM

SCIENCE PRIORITIES

Who Will Pay for Public Access to Research Data?

When economic models and infrastructure are not in place to ensure access and preservation, federally funded research data are "at risk."

Francine Berman[1] and Vint Cerf[2]

On 22 February, the U.S. Office of Science and Technology Policy (OSTP) released a memo calling for public access for publications and data resulting from federally sponsored research grants (*1*). The memo directed federal agencies with more than $100 million R&D expenditures to "develop a plan to support increased public access to the results of research funded by the Federal Government." Perhaps even more succinctly, a subsequent *New York Times* opinion page sported the headline "We Paid for the Research, So Let's See It" (*2*). So who pays for data infrastructure?

What happens to valuable data when project funding ends? Consider, for example, a 3-year research project in which valuable sensor data are collected from an environmentally sensitive area. Those data may be useful not just for the duration of the project but for the next decade or more to collaborators and a broader community of researchers. For the first 3 years, the costs of stewardship (including development of a database that supports analysis, access to the

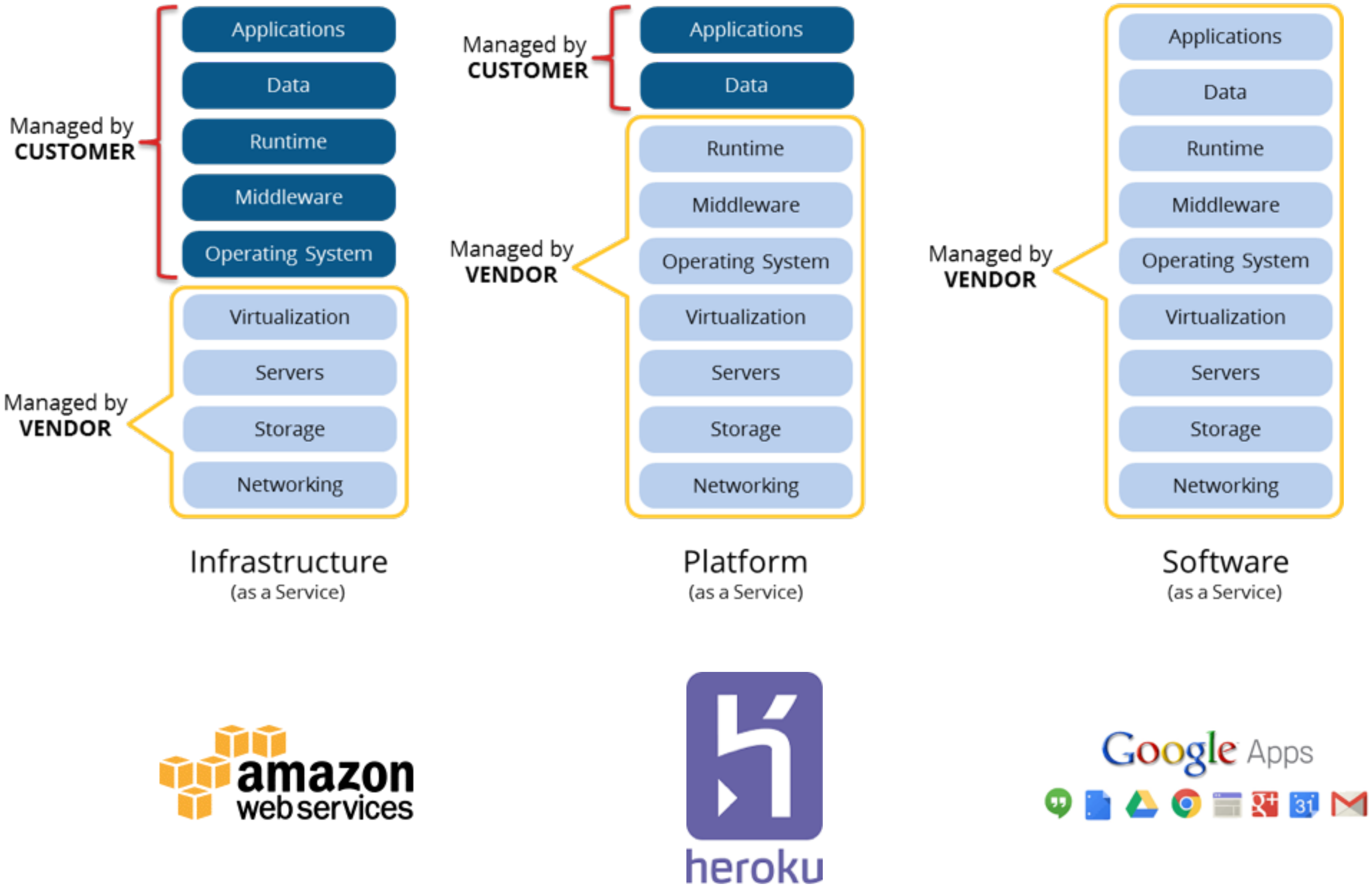
Before we decide who is paying, lets figure out the bill.

What are the costs of running a open data, and curation programs?

- Infrastructure (hardware, software & implementation)
- Governance
- Staff & Skills Development
- Maintenance

Infrastructure

- Open data portal development costs
    - (Open source is free like a pony)
- Cost of storage - including cloud storage, tape backups, and local storage
- Cost of publishing datasets, creating APIs, and automating preservation actions.
- Hardware (rent or roll your own).
    - X_as_a_Service
    - SaaS, Paas, IaaS



Managed by **CUSTOMER**

Applications

Data

Runtime

Middleware

Operating System

Managed by **VENDOR**

Virtualization

Servers

Storage

Networking

Infrastructure
(as a Service)

Managed by **CUSTOMER**

Applications

Data

Managed by **VENDOR**

Runtime

Middleware

Operating System

Virtualization

Servers

Storage

Networking

Platform
(as a Service)

Managed by **VENDOR**

Applications

Data

Runtime

Middleware

Operating System

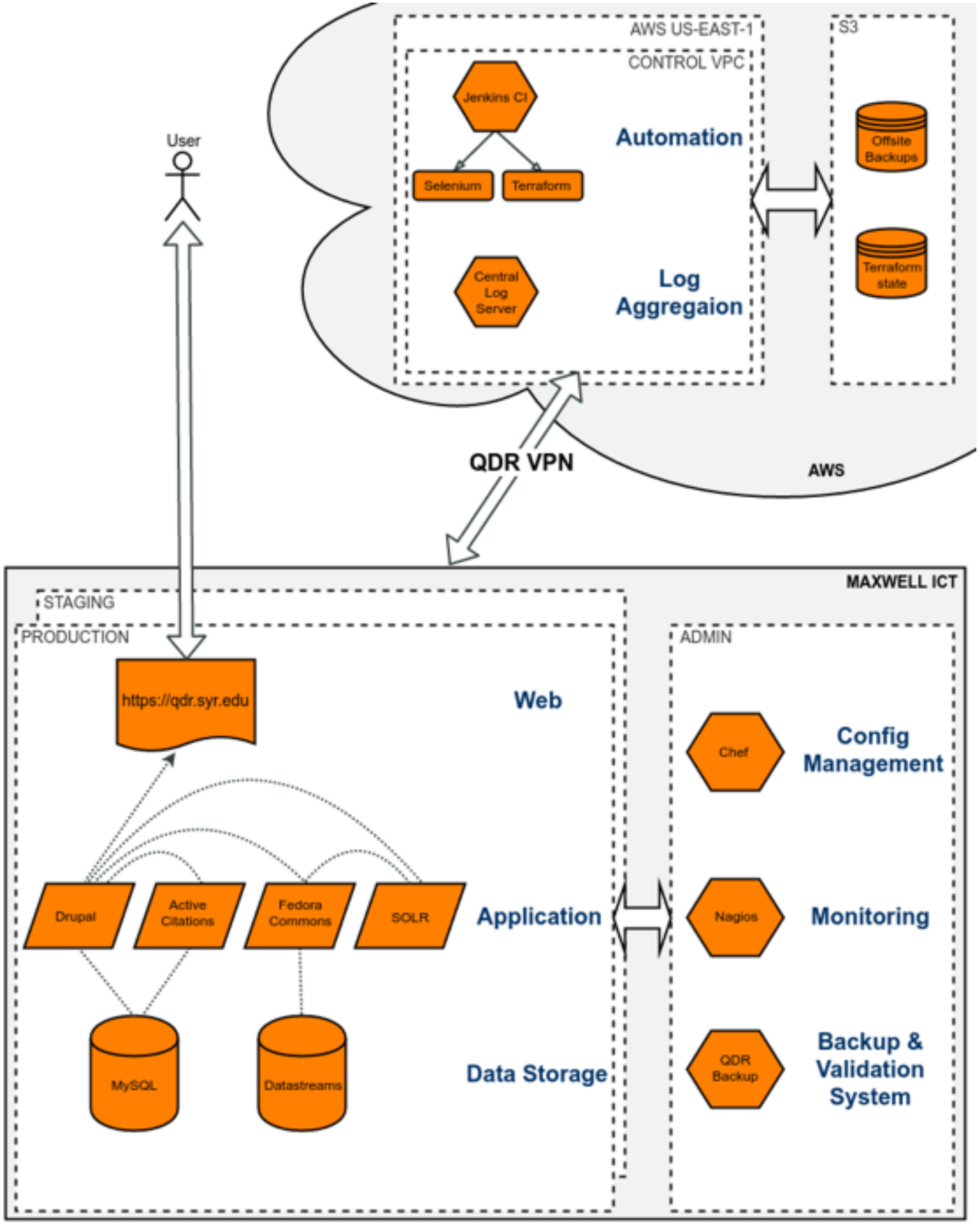
Virtualization

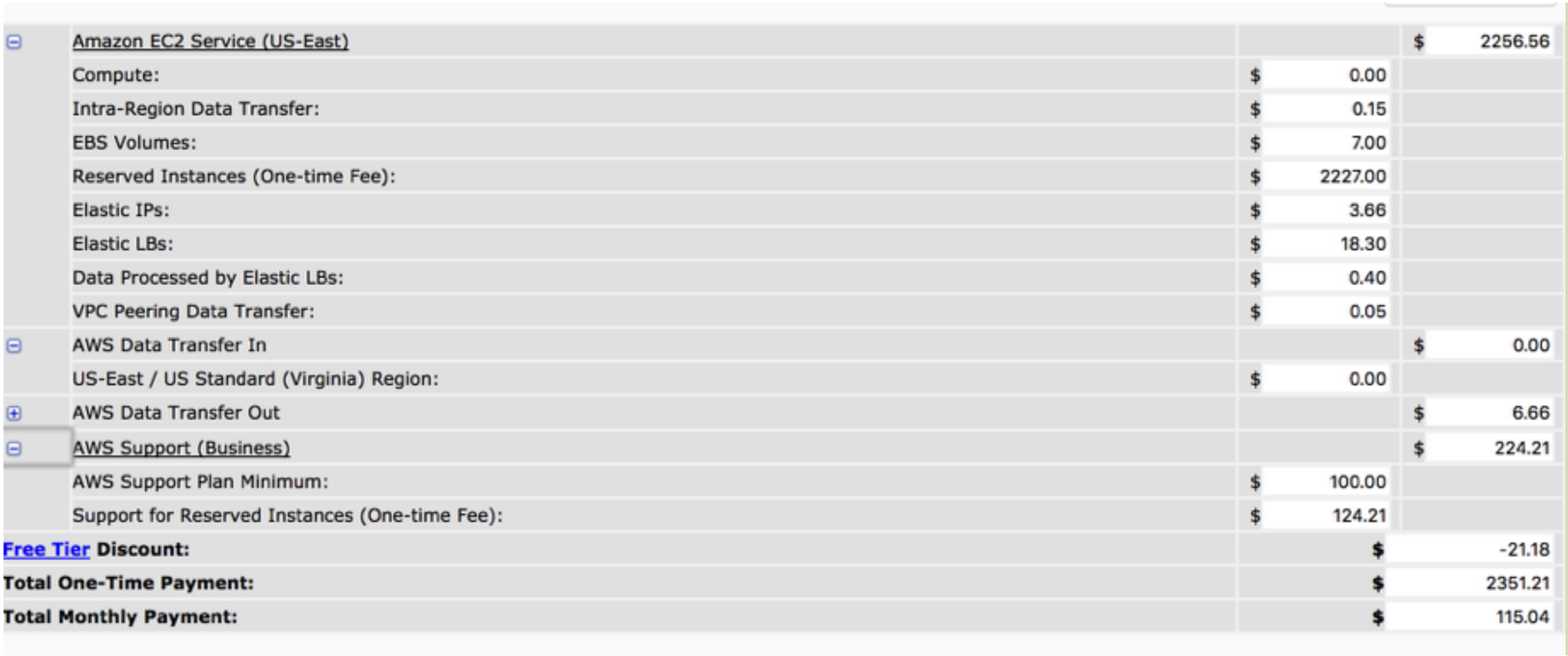
Servers

Storage

Networking

Software
(as a Service)


AWS US-EAST-1
CONTROL VPC
Jenkins CI
Selenium
Terraform
Automation
Central Log Server
Log Aggregaion
S3
Offsite Backups
Terraform state
AWS
User
QDR VPN
MAXWELL ICT
STAGING
PRODUCTION
https://qdr.syr.edu
Web
Drupal
Active Citations
Fedora Commons
SOLR
Application
MySQL
Datastreams
Data Storage
ADMIN
Chef
Config Management
Nagios
Monitoring
QDR Backup
Backup & Validation System



⊟	Amazon EC2 Service (US-East)			$	2256.56
	Compute:	$	0.00		
	Intra-Region Data Transfer:	$	0.15		
	EBS Volumes:	$	7.00		
	Reserved Instances (One-time Fee):	$	2227.00		
	Elastic IPs:	$	3.66		
	Elastic LBs:	$	18.30		
	Data Processed by Elastic LBs:	$	0.40		
	VPC Peering Data Transfer:	$	0.05		
⊟	AWS Data Transfer In			$	0.00
	US-East / US Standard (Virginia) Region:	$	0.00		
⊞	AWS Data Transfer Out			$	6.66
⊟	AWS Support (Business)			$	224.21
	AWS Support Plan Minimum:	$	100.00		
	Support for Reserved Instances (One-time Fee):	$	124.21		
Free Tier Discount:				**$**	-21.18
Total One-Time Payment:				**$**	2351.21
Total Monthly Payment:				**$**	115.04

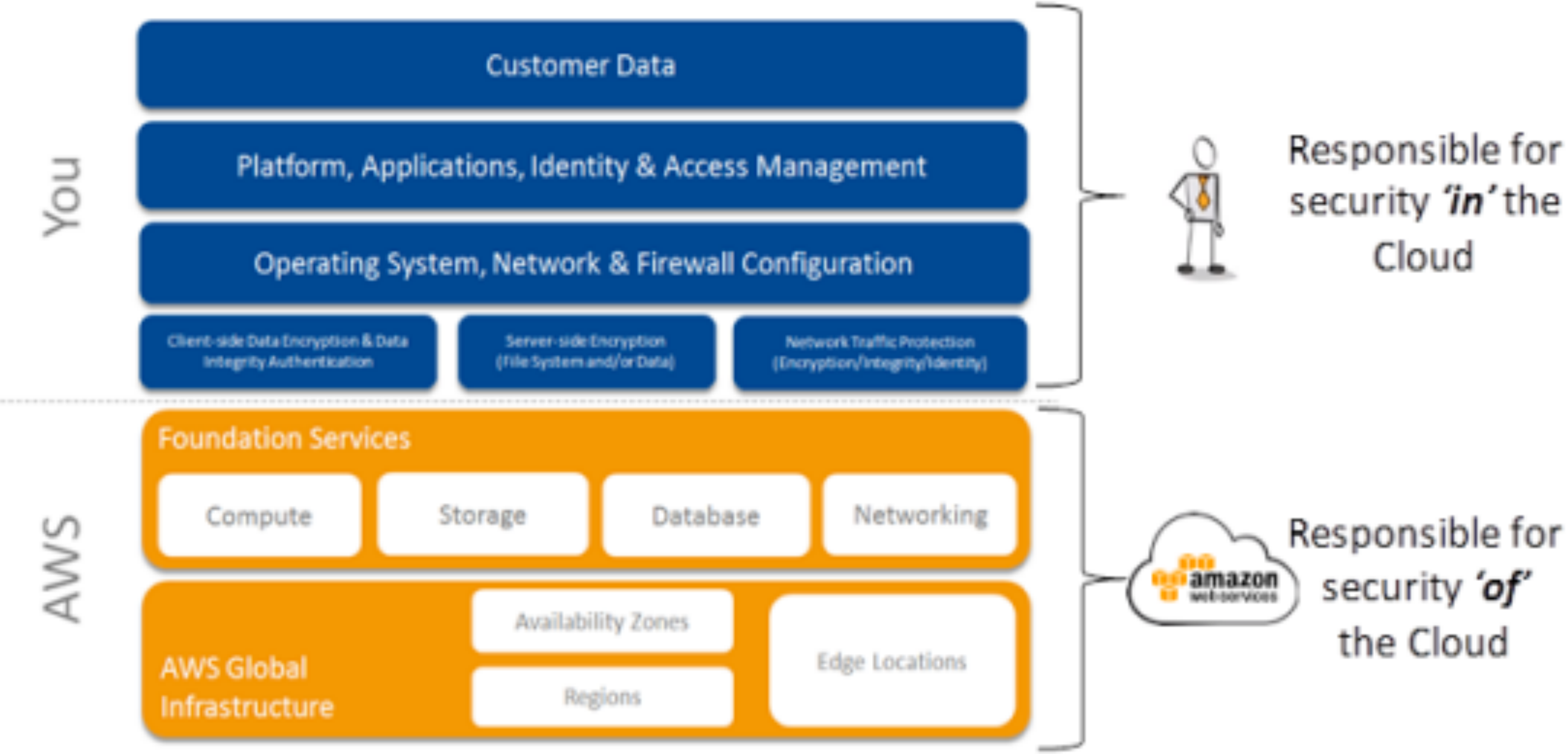
Total = $338.00 Month average

Total infrastructure costs are ~ $10k per year.

Governance

- Getting legislation or institutional policy written and approved
    - Maintaining and enforcing the policy or legal requirements
    - Legal costs to comply with open data legislation
- Managing requests or questions about datasets
- Protecting privacy / providing security

CSA cloud security alliance®



You

Customer Data

Platform, Applications, Identity & Access Management

Operating System, Network & Firewall Configuration

Client-side Data Encryption & Data Integrity Authentication

Server-side Encryption (File System and/or Data)

Network Traffic Protection (Encryption/Integrity/Identity)

Responsible for security *'in'* the Cloud

AWS

Foundation Services

Compute

Storage

Database

Networking

AWS Global Infrastructure

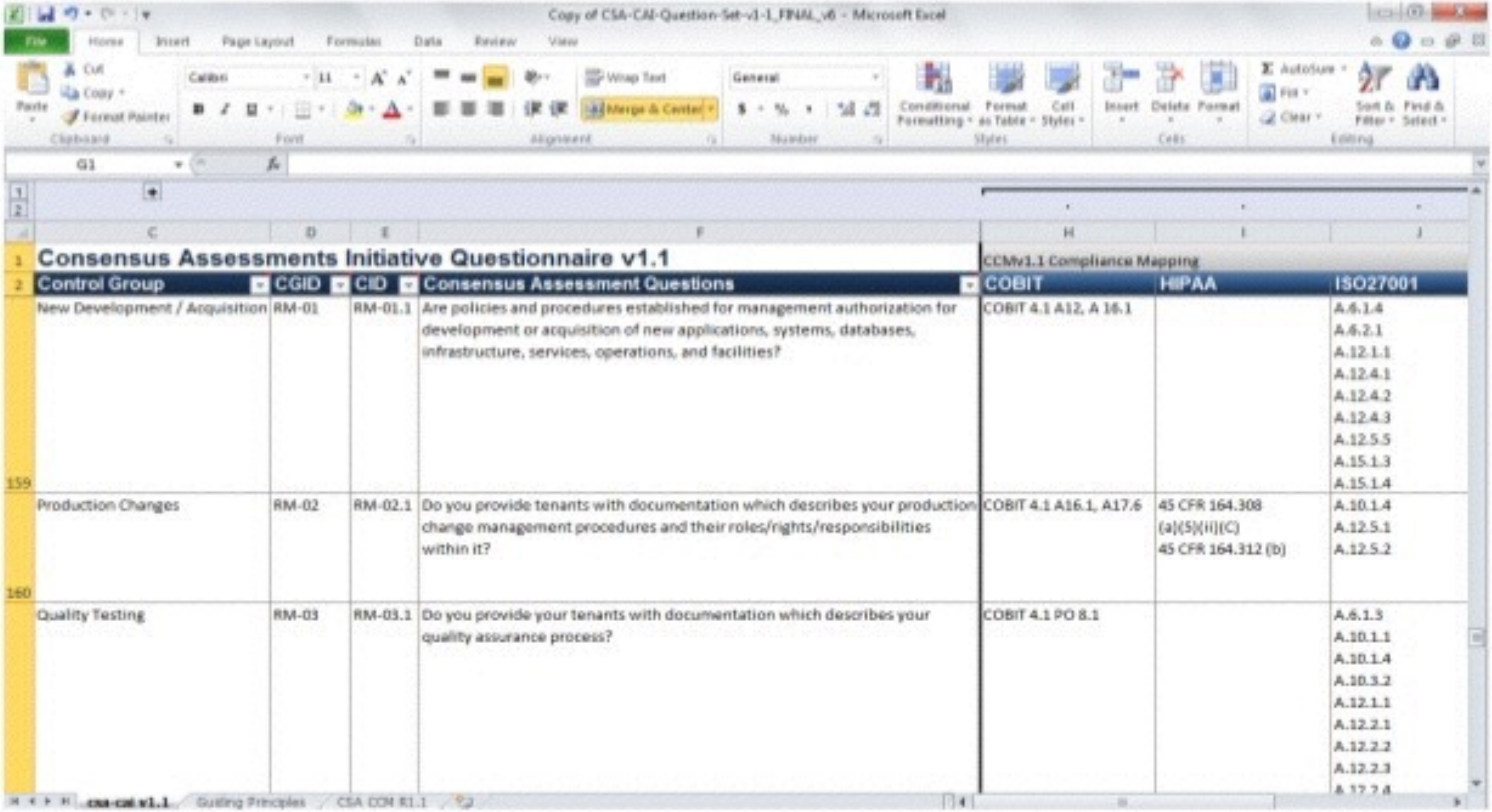
Availability Zones

Regions

Edge Locations

amazon web services

Responsible for security *'of'* the Cloud



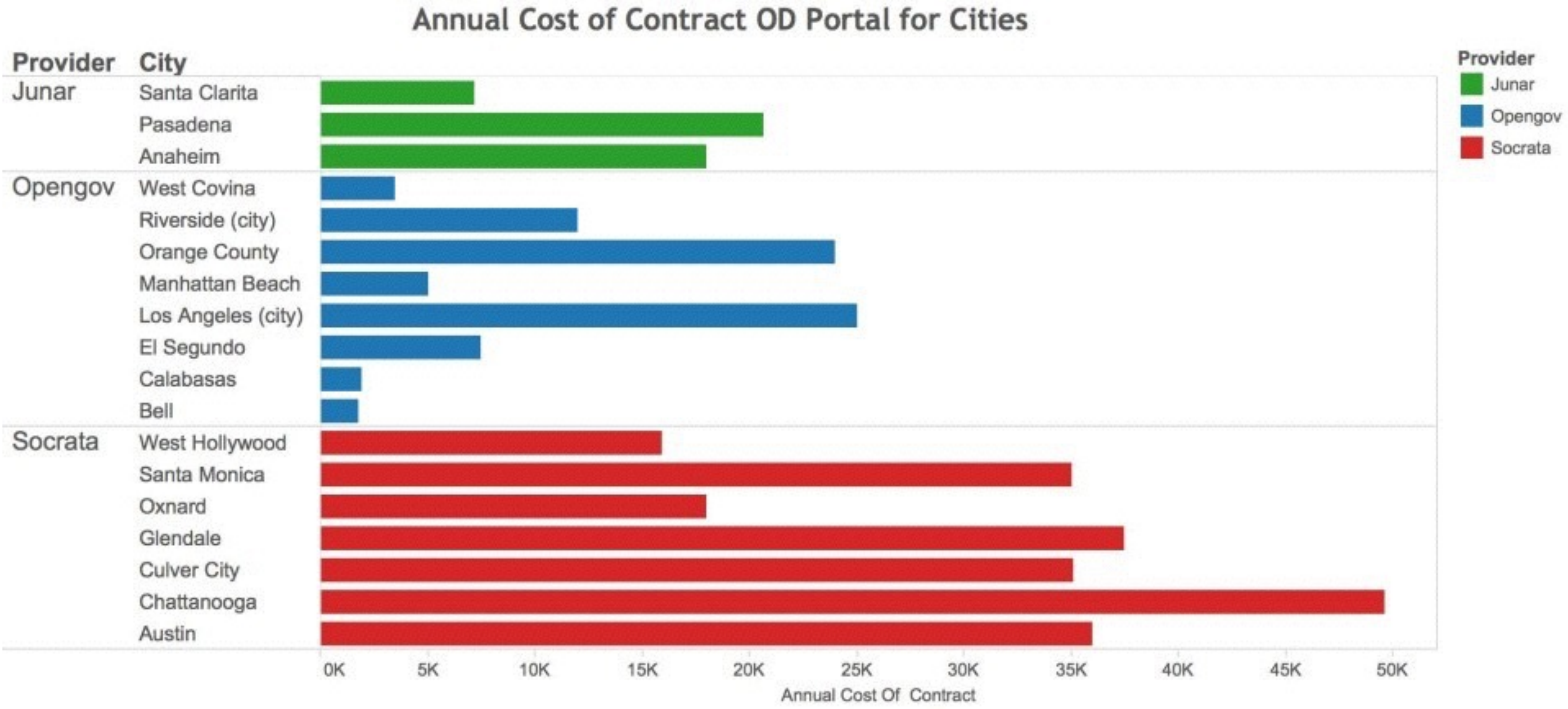
1 Consensus Assessments Initiative Questionnaire v1.1 | CCMv1.1 Compliance Mapping

2	Control Group	CGID	CID	Consensus Assessment Questions	COBIT	HIPAA	ISO27001
159	New Development / Acquisition	RM-01	RM-01.1	Are policies and procedures established for management authorization for development or acquisition of new applications, systems, databases, infrastructure, services, operations, and facilities?	COBIT 4.1 A12, A 16.1		A.6.1.4 A.6.2.1 A.12.1.1 A.12.4.1 A.12.4.2 A.12.4.3 A.12.5.5 A.15.1.3 A.15.1.4
160	Production Changes	RM-02	RM-02.1	Do you provide tenants with documentation which describes your production change management procedures and their roles/rights/responsibilities within it?	COBIT 4.1 A16.1, A17.6	45 CFR 164.308 (a)(5)(ii)(C) 45 CFR 164.312 (b)	A.10.1.4 A.12.5.1 A.12.5.2
	Quality Testing	RM-03	RM-03.1	Do you provide your tenants with documentation which describes your quality assurance process?	COBIT 4.1 PO 8.1		A.6.1.3 A.10.1.1 A.10.1.4 A.10.3.2 A.12.1.1 A.12.2.1 A.12.2.2 A.12.2.3 A.12.2.4

https://figshare.com/articles/Qualitative_Data_Repository_Consensus_Assessment_Initiative_Questionnaire_for_AWS/4670506

Sustainability

- Hardware and software maintenance cost
- Data inventorying, publishing, optimizing discovery, updating relevant standards, preservation checks / actions
- Labor for doing both of the above tasks
- Liability costs in case of publication of nonpublic information (i.e. have a lawyer on retainer)
- Impact and analytics
- Vendor contracts (EzID; Socrate; perma.io; etc)



https://govex.jhu.edu/open-data-how-much-does-it-cost/

Staff & Skills Development

- Developers (front and back end), and curators
- Awareness raising activities to promote use
- Capacity building for the use of data within government, discipline, or school
- Ongoing training and capacity building

QDR

- Directors
    - Content - full time (salary)
    - Technical - .05 time (salary)
- Curators - 3 part-time (GA student salary)
- Developers
    - Front-end - Contract - $75/ hr
    - System Admin / Development / Operations - $70/hr

Labor costs > $110,000 / annual.

UKDA

- employs 64.5 people.
- total budget UK Data Archive (2010-11) ~ £3.43 million
    - £2.43 million - Staff
    - £1 million - Infrastructure

Dryad

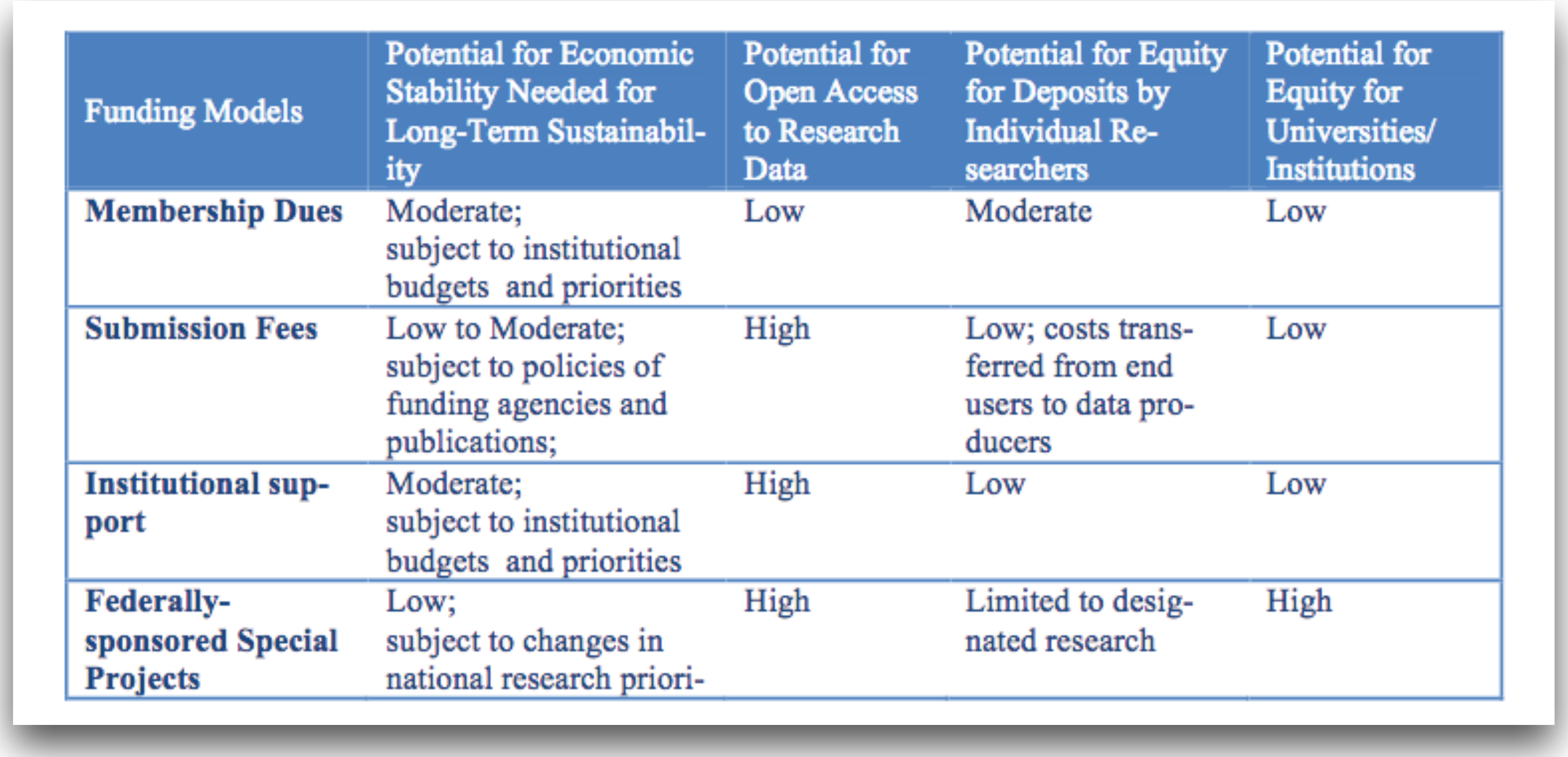
- 4-6 FTE
- Total budget $350,000 per year
    - $300,000 - Staff
    - $5,000-$10,000 - Infrastructure

Cost Recovery
(academic, discipline or research repositories)

1. **Membership**
2. **Submission fees**

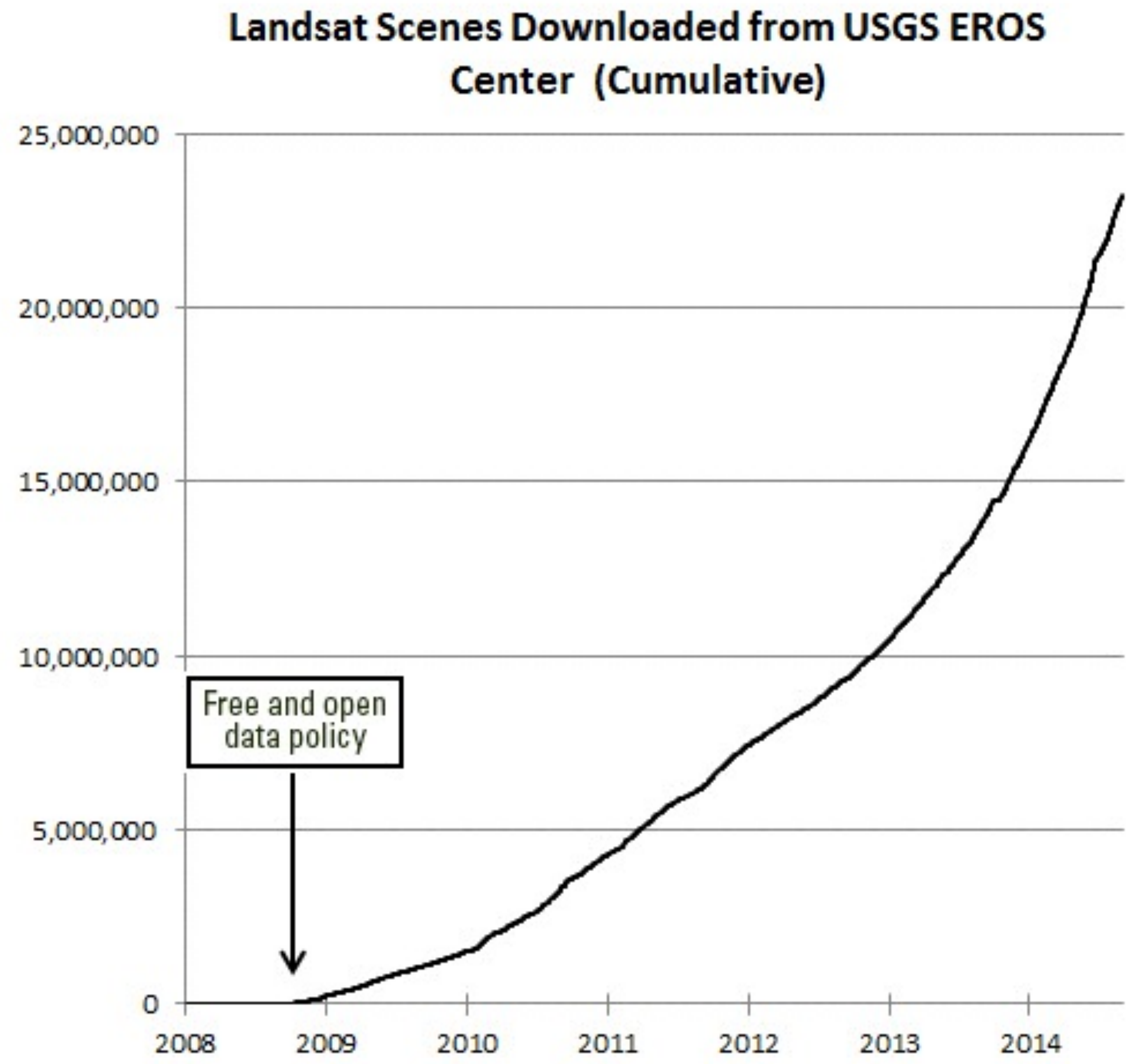
Both of these create a barrier to entry, they differ on assumes the cost.

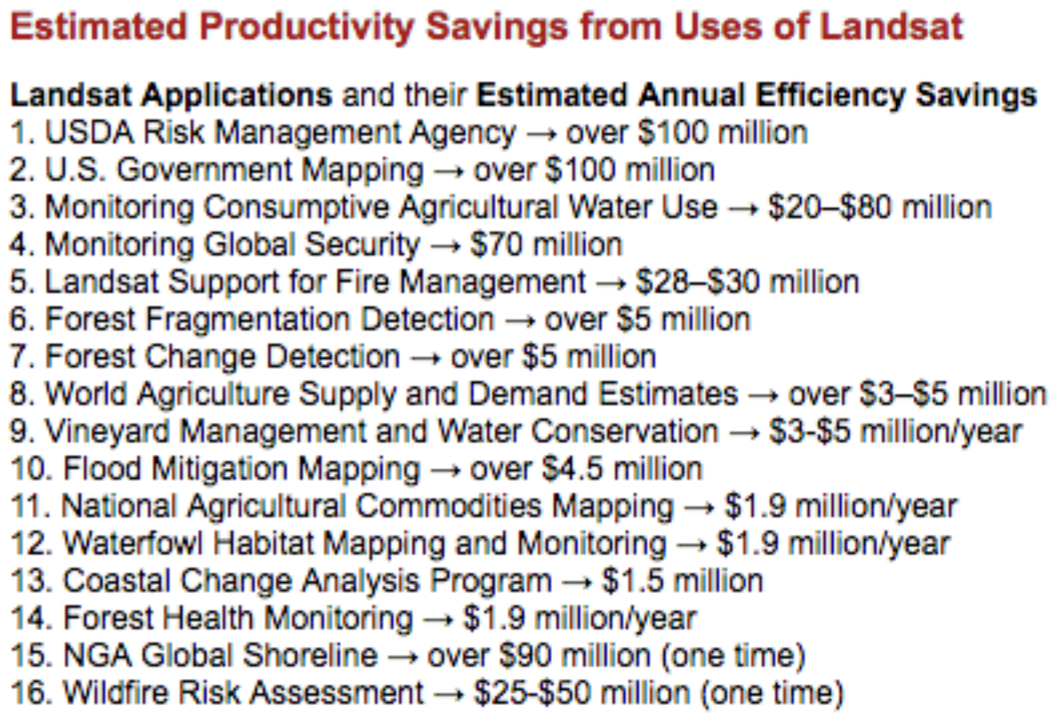
3. **Institutional Support**
4. **Federal funding for Special Projects**



Funding Models	Potential for Economic Stability Needed for Long-Term Sustainability	Potential for Open Access to Research Data	Potential for Equity for Deposits by Individual Researchers	Potential for Equity for Universities/ Institutions
Membership Dues	Moderate; subject to institutional budgets and priorities	Low	Moderate	Low
Submission Fees	Low to Moderate; subject to policies of funding agencies and publications;	High	Low; costs transferred from end users to data producers	Low
Institutional support	Moderate; subject to institutional budgets and priorities	High	Low	Low
Federally-sponsored Special Projects	Low; subject to changes in national research priori-	High	Limited to designated research	High

http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf



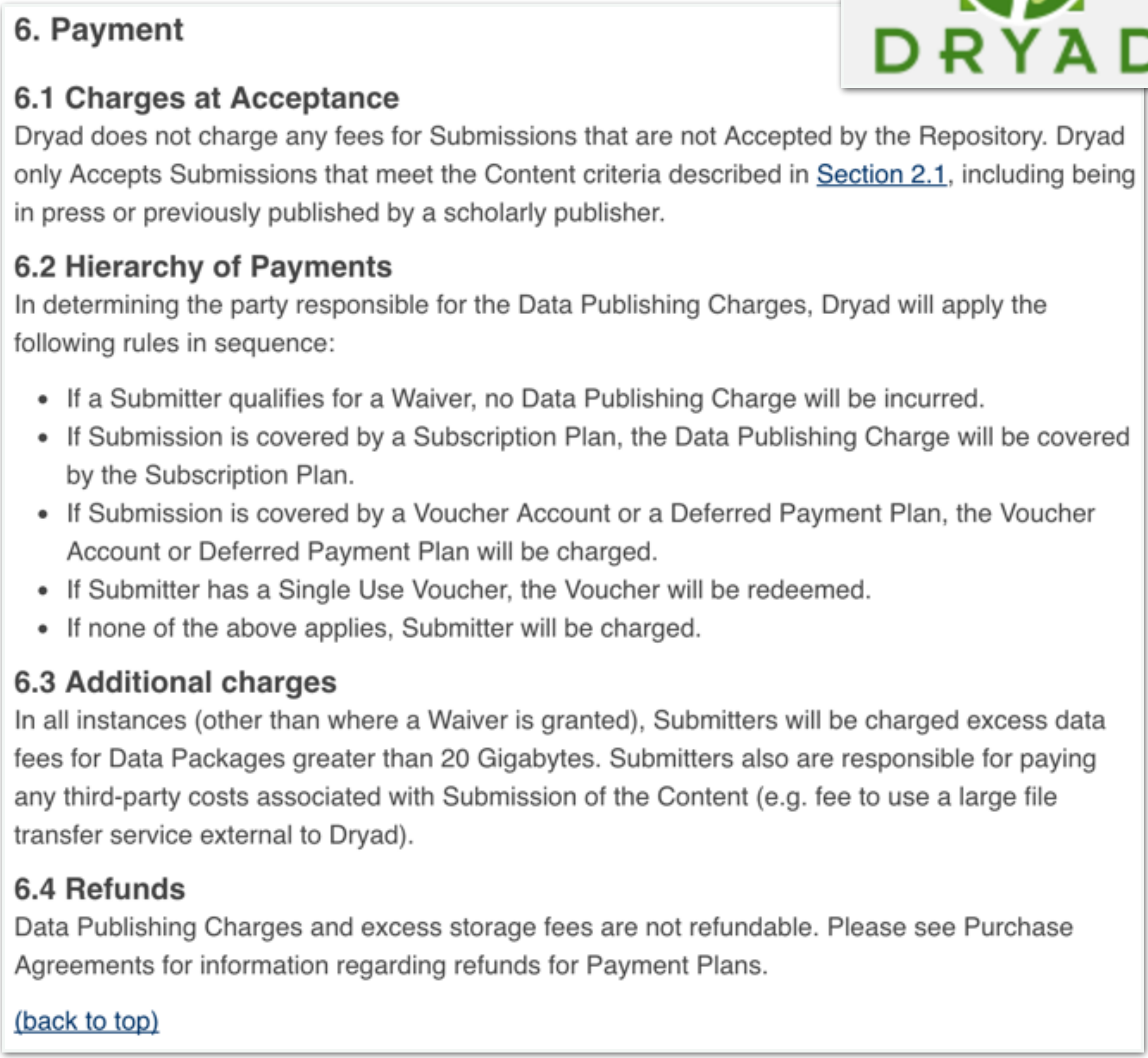


Estimated Productivity Savings from Uses of Landsat

Landsat Applications and their **Estimated Annual Efficiency Savings**
1. USDA Risk Management Agency → over $100 million
2. U.S. Government Mapping → over $100 million
3. Monitoring Consumptive Agricultural Water Use → $20–$80 million
4. Monitoring Global Security → $70 million
5. Landsat Support for Fire Management → $28–$30 million
6. Forest Fragmentation Detection → over $5 million
7. Forest Change Detection → over $5 million
8. World Agriculture Supply and Demand Estimates → over $3–$5 million
9. Vineyard Management and Water Conservation → $3-$5 million/year
10. Flood Mitigation Mapping → over $4.5 million
11. National Agricultural Commodities Mapping → $1.9 million/year
12. Waterfowl Habitat Mapping and Monitoring → $1.9 million/year
13. Coastal Change Analysis Program → $1.5 million
14. Forest Health Monitoring → $1.9 million/year
15. NGA Global Shoreline → over $90 million (one time)
16. Wildfire Risk Assessment → $25-$50 million (one time)

$436 million per year

https://landsat.gsfc.nasa.gov/landsat-seen-as-stunning-return-on-public-investment/



6. Payment

6.1 Charges at Acceptance

Dryad does not charge any fees for Submissions that are not Accepted by the Repository. Dryad only Accepts Submissions that meet the Content criteria described in Section 2.1, including being in press or previously published by a scholarly publisher.

6.2 Hierarchy of Payments

In determining the party responsible for the Data Publishing Charges, Dryad will apply the following rules in sequence:

- If a Submitter qualifies for a Waiver, no Data Publishing Charge will be incurred.
- If Submission is covered by a Subscription Plan, the Data Publishing Charge will be covered by the Subscription Plan.
- If Submission is covered by a Voucher Account or a Deferred Payment Plan, the Voucher Account or Deferred Payment Plan will be charged.
- If Submitter has a Single Use Voucher, the Voucher will be redeemed.
- If none of the above applies, Submitter will be charged.

6.3 Additional charges

In all instances (other than where a Waiver is granted), Submitters will be charged excess data fees for Data Packages greater than 20 Gigabytes. Submitters also are responsible for paying any third-party costs associated with Submission of the Content (e.g. fee to use a large file transfer service external to Dryad).

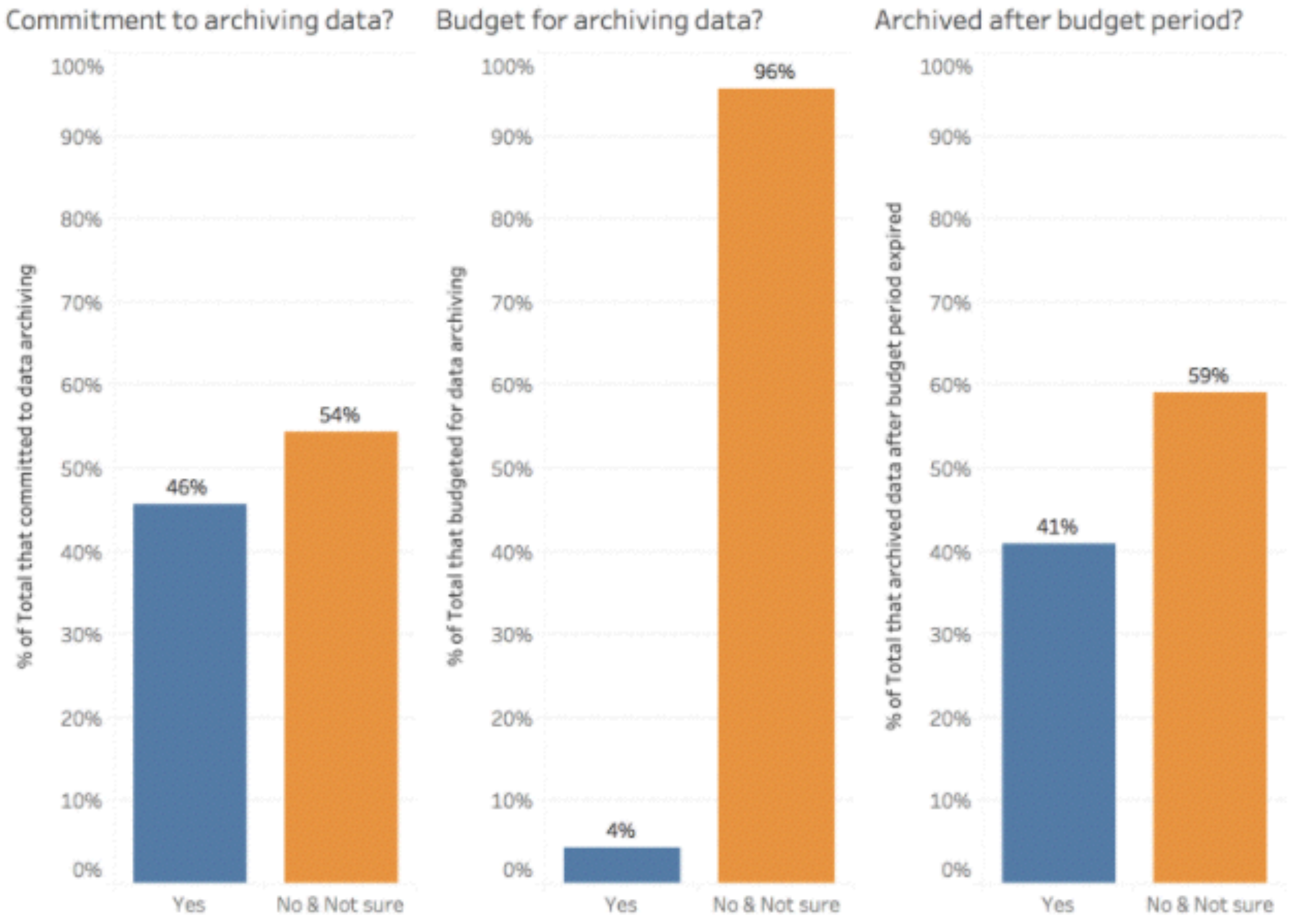
6.4 Refunds

Data Publishing Charges and excess storage fees are not refundable. Please see Purchase Agreements for information regarding refunds for Payment Plans.

(back to top)

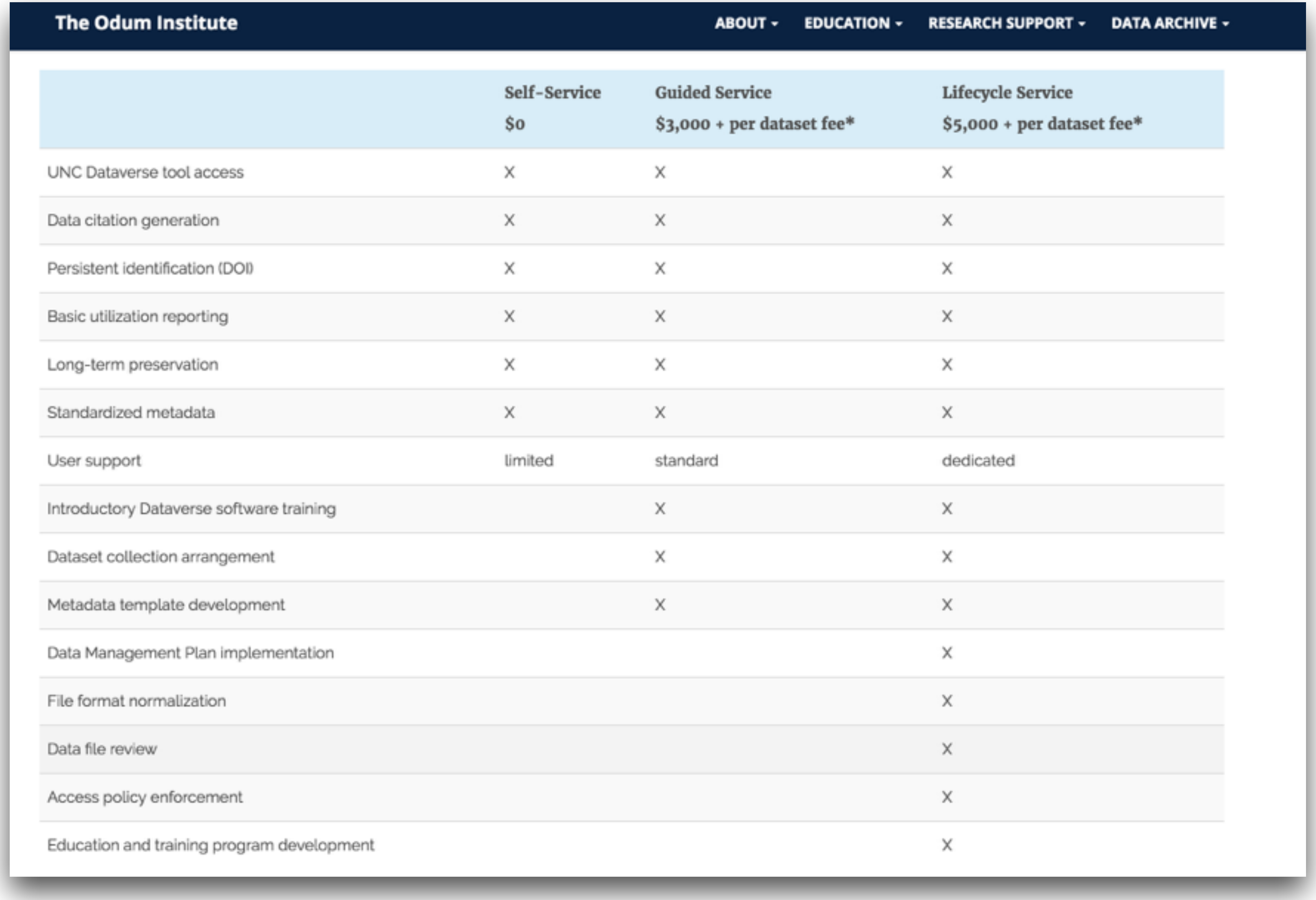
DRYAD

$120 per archived dataset



https://blog.datadryad.org/2017/04/18/how-do-researchers-pay-for-data-publishing-results-of-a-recent-submitter-survey/

- Back of the envelope math:
    - 2016 budget = $350,000
    - Cost for data deposit $120
    - Total data deposited 4258
        - Minus waiver rate of 1/3 of all submissions
    - Total 2016 revenue = $357,720. ($7000 margin!)



	Self-Service $0	Guided Service $3,000 + per dataset fee*	Lifecycle Service $5,000 + per dataset fee*
UNC Dataverse tool access	X	X	X
Data citation generation	X	X	X
Persistent identification (DOI)	X	X	X
Basic utilization reporting	X	X	X
Long-term preservation	X	X	X
Standardized metadata	X	X	X
User support	limited	standard	dedicated
Introductory Dataverse software training		X	X
Dataset collection arrangement		X	X
Metadata template development		X	X
Data Management Plan implementation			X
File format normalization			X
Data file review			X
Access policy enforcement			X
Education and training program development			X