NIST Challenges in Materials Science, Data, and Repositories

John Henry J. Scott
Physicist

James A. Warren
Technical Program Director for Materials Genomics

Material Measurement Laboratory
National Institute of Standards and Technology

Gaithersburg, MD

The Materials Genome Initiative (MGI)

Goal

decrease the
time-to-market
and
cost
for new materials
by >50%

THE U.S. MATERIALS GENOME INITIATIVE

"...to discover, develop, and deploy new materials twice as fast, we're launching what we call the Materials Genome Initiative"
– President Obama, 2011

Meeting Societal Needs

Advanced materials are at the heart of innovation, economic opportunities, and global competitiveness. They are the foundation for new capabilities, tools, and technologies that meet urgent societal needs including clean energy, human welfare, and national security.

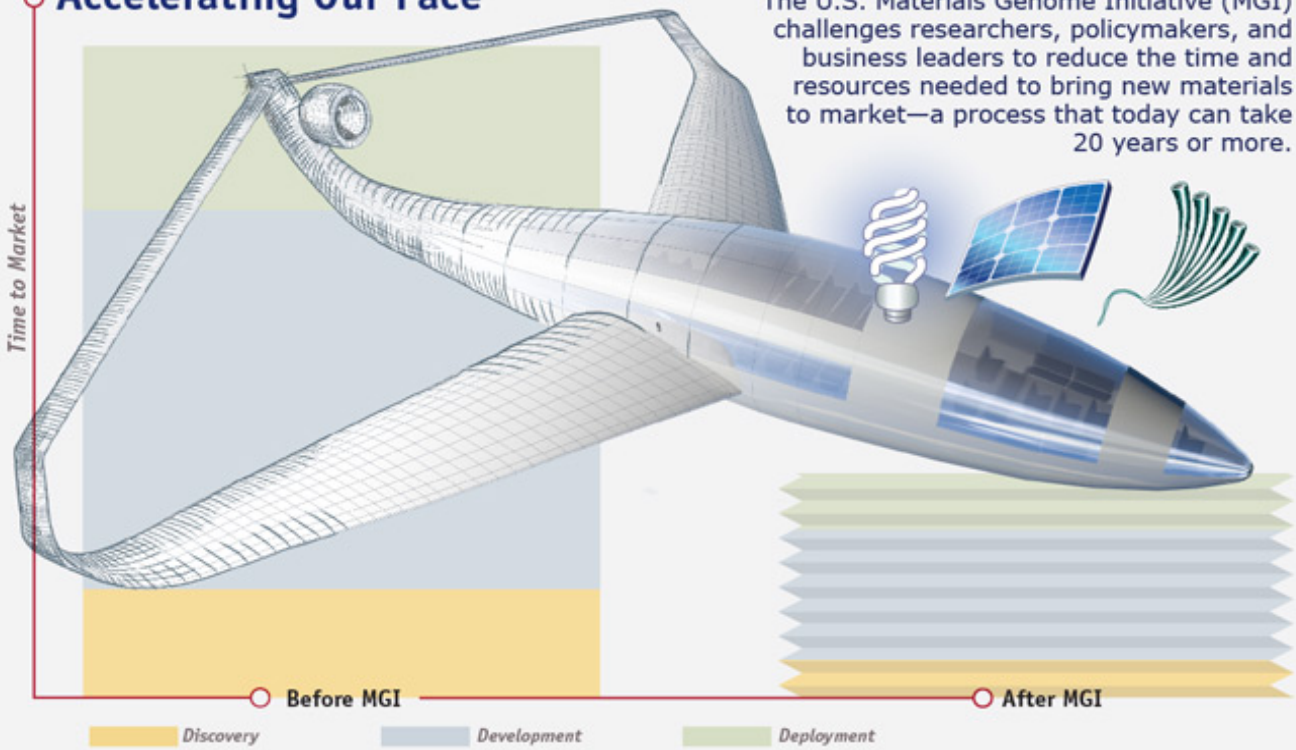
Clean Energy

Human Welfare

National Security

Accelerating Our Pace

The U.S. Materials Genome Initiative (MGI) challenges researchers, policymakers, and business leaders to reduce the time and resources needed to bring new materials to market—a process that today can take 20 years or more.



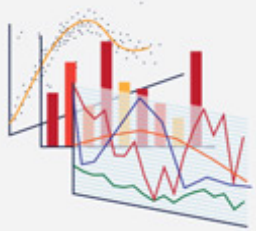
Building Infrastructure for Success

The MGI is a multi-agency initiative to renew investments in infrastructure designed for performance, and to foster a more open, collaborative approach to developing advanced materials, helping U.S. Institutions accelerate their time-to-market.

Computational tools

Experimental tools

Collaborative networks



Digital data

THE U.S. MATERIALS GENOME INITIATIVE

"...to discover, develop, and deploy new materials twice as fast, we're launching what we call the Materials Genome Initiative"
– President Obama, 2011

Meeting Societal Needs

Advanced materials are at the heart of innovation, economic opportunities, and global competitiveness. They are the foundation for new capabilities, tools, and technologies that meet urgent societal needs including clean energy, human welfare, and national security.

Clean Energy

Human Welfare

National Security

Accelerating Our Pace

The U.S. Materials Genome Initiative (MGI) challenges researchers, policymakers, and business leaders to reduce the time and resources needed to bring new materials to market—a process that today can take 20 years or more.

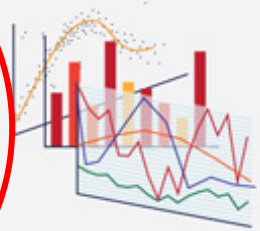
Building Infrastructure for Success

The MGI is a multi-agency initiative to renew investments in infrastructure designed for performance, and to foster a more open, collaborative approach to developing advanced materials, helping U.S. Institutions accelerate their time-to-market.

Computational tools

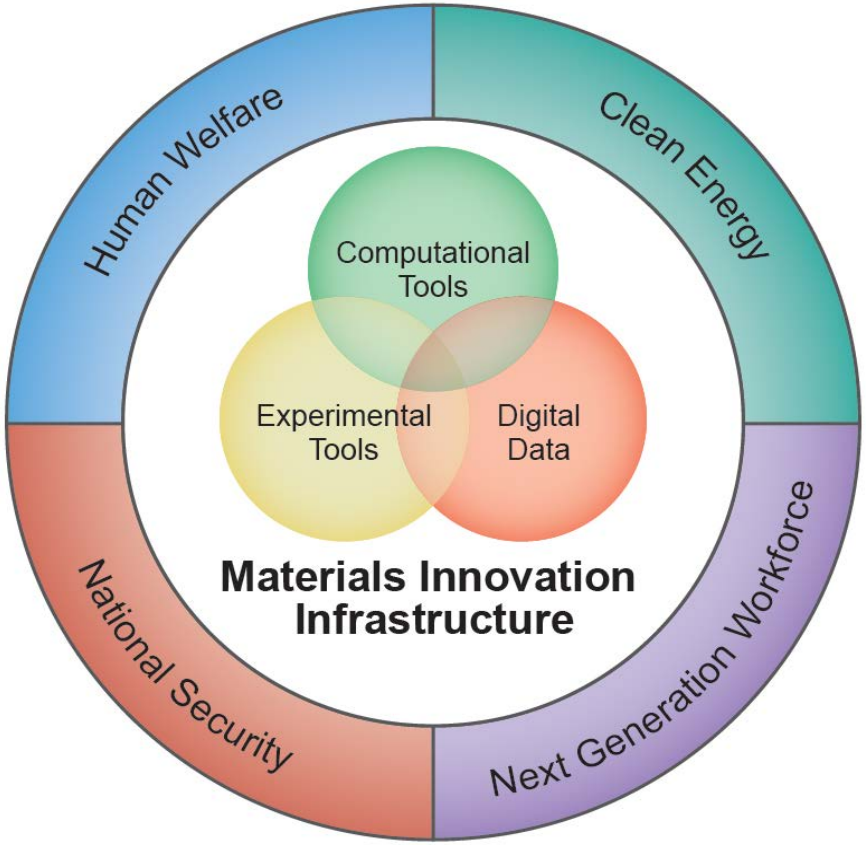
Experimental tools

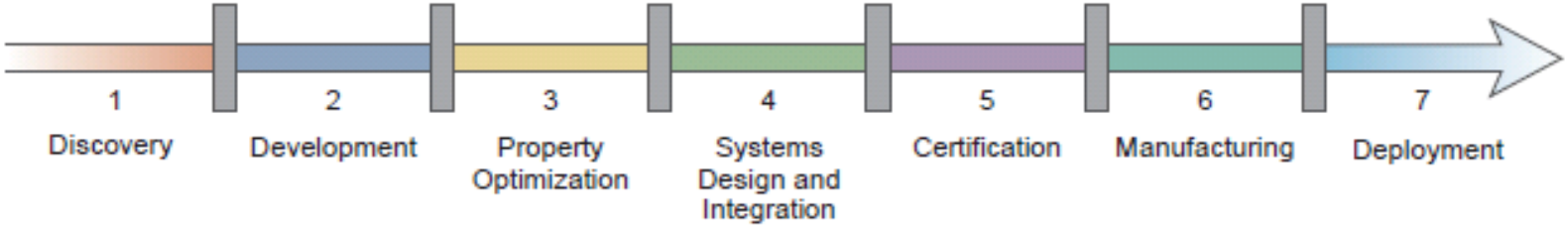
Collaborative networks

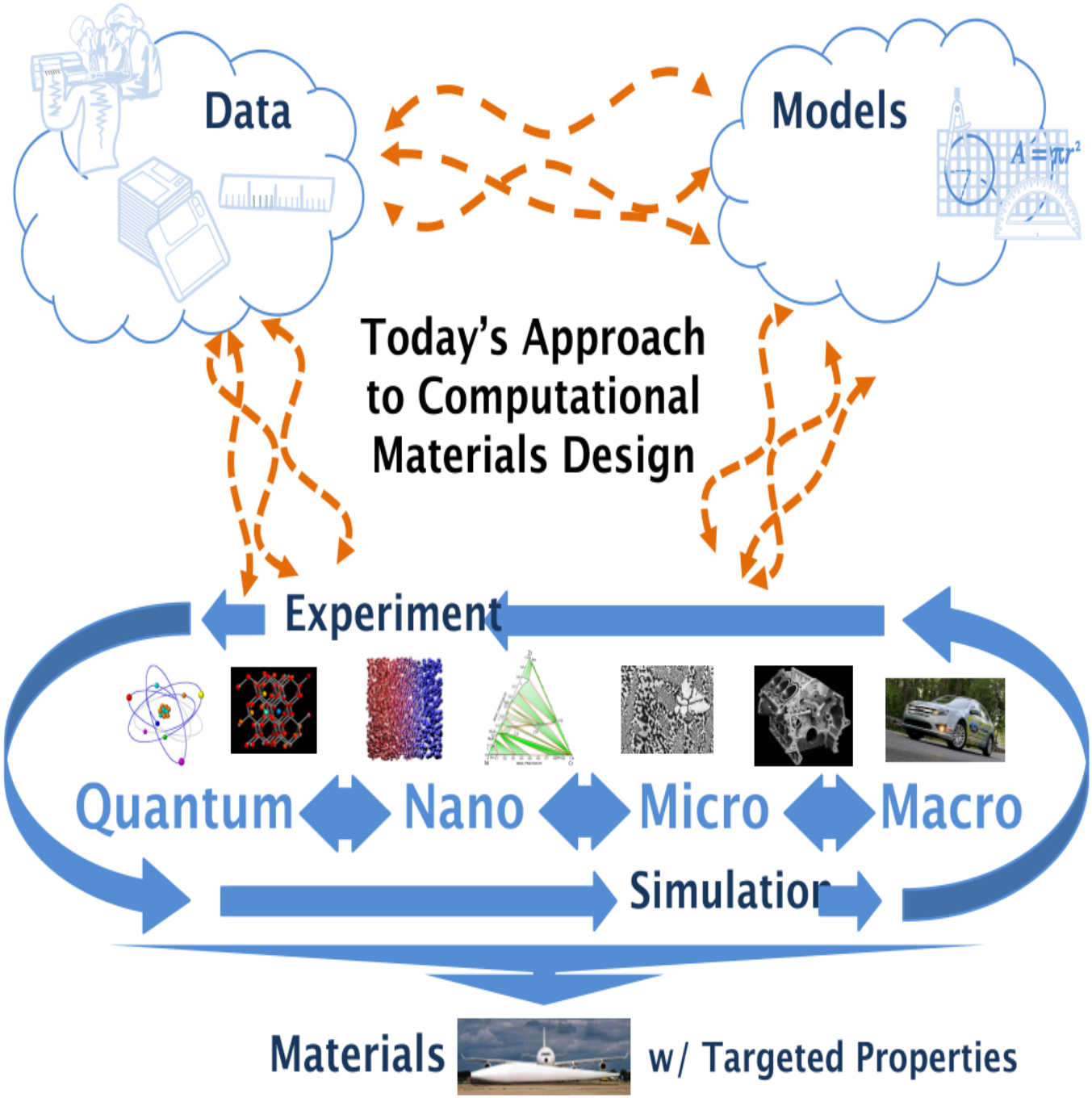


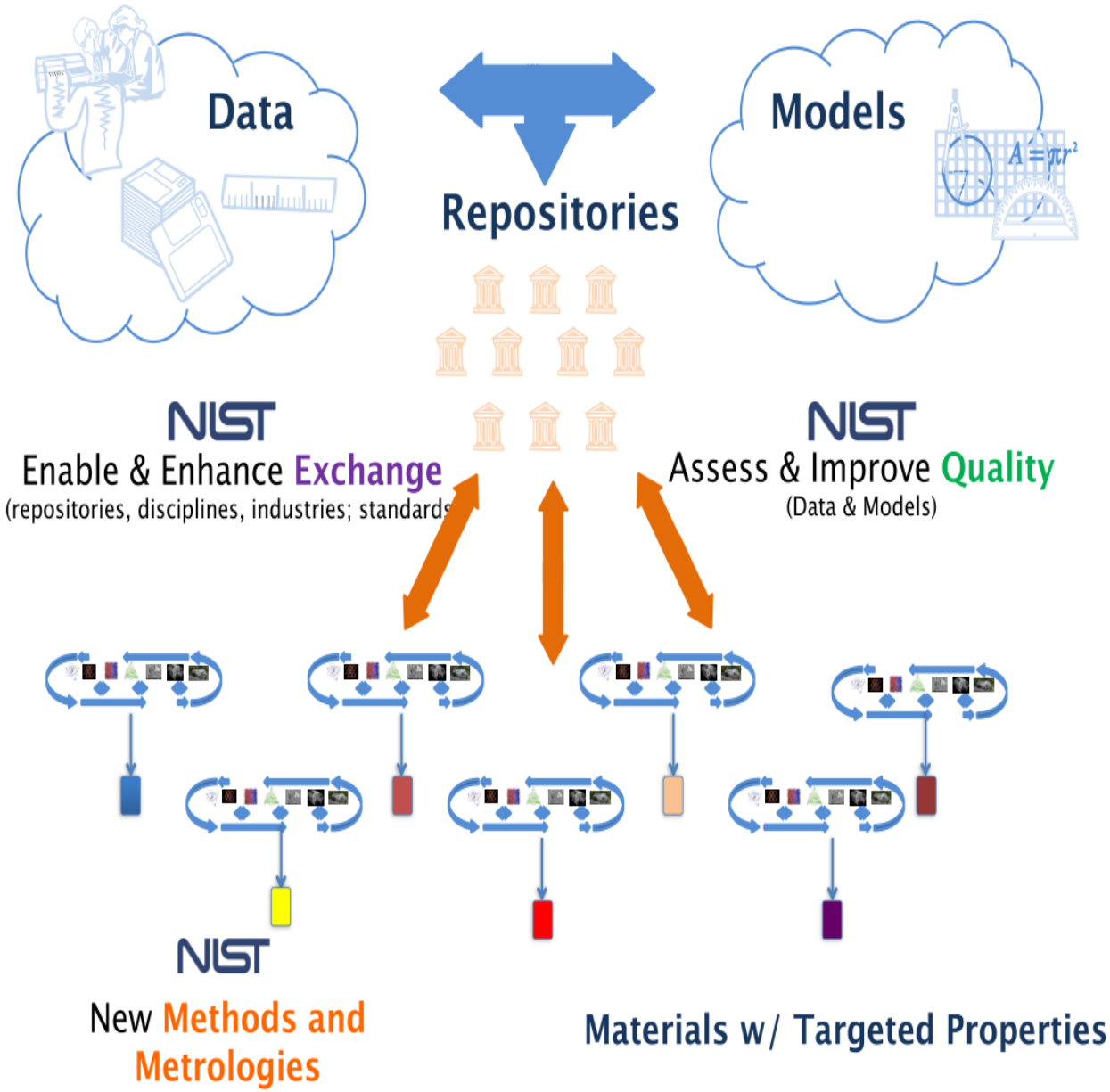
Digital data

The Materials Innovation Infrastructure






Data
Models
Today's Approach to Computational Materials Design
Experiment
Quantum
Nano
Micro
Macro
Simulation
Materials
w/ Targeted Properties


Data
Repositories
Models
$A = \pi r^2$
NIST
Enable & Enhance Exchange
(repositories, disciplines, industries; standards
NIST
Assess & Improve Quality
(Data & Models)
NIST
New Methods and Metrologies
Materials w/ Targeted Properties

Scope: Goals of the Initiative

Goal 1: NIST establishes essential materials data and model exchange protocols

Goal 2: NIST establishes the means to ensure the quality of materials data and models

Goal 3: NIST establishes new methods, metrologies and capabilities necessary for accelerated materials development

Enable and Enhance Exchange

- Develop and deploy repositories
- Develop and disseminate materials informatics infrastructure
  - Enable data discovery through tools and standards
  - Capture data from scientific workflows and archival sources
  - Engage with stakeholders to determine needs and disseminate best practices
- Integrate across length and time scale
- Build and Test infrastructure through Pilots

Data Tools and Informatics for Materials Data

Need

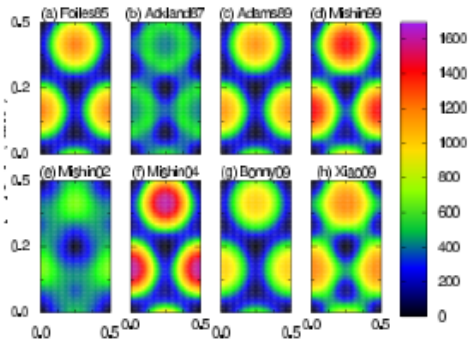
- Improved efficiency and reproducibility of thermodynamic and kinetic simulations (e.g. CALPHAD, first principles, atomistic methods).
- Accessible phase-based data described as functions of composition, temperature and pressure.

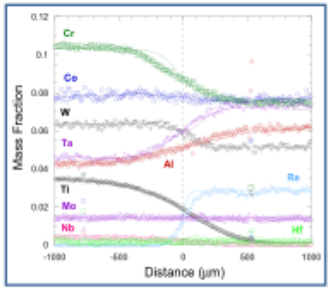
Objectives

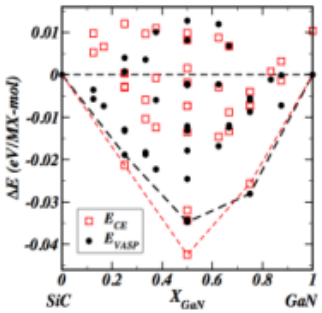
- Develop file repositories that enable links to data files and the ability to identify key metadata.
- Develop informatics tools to enable data capture and retrieval of phase-based data.
- Develop tools that are available for the community to contribute towards and use.

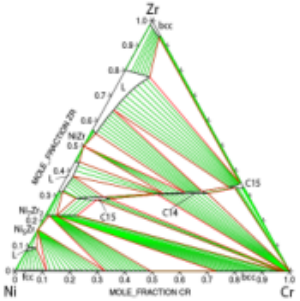
Achievements and Impact

- NIST D-Space File Repository established with Kent State (http://nist.matdl.org/)
- NIST Interatomic Potentials Repository (www.ctcms.nist.gov/potentials) : *100+ down-loadable referenced interatomic potentials*
- Materials Data Curator system is under development





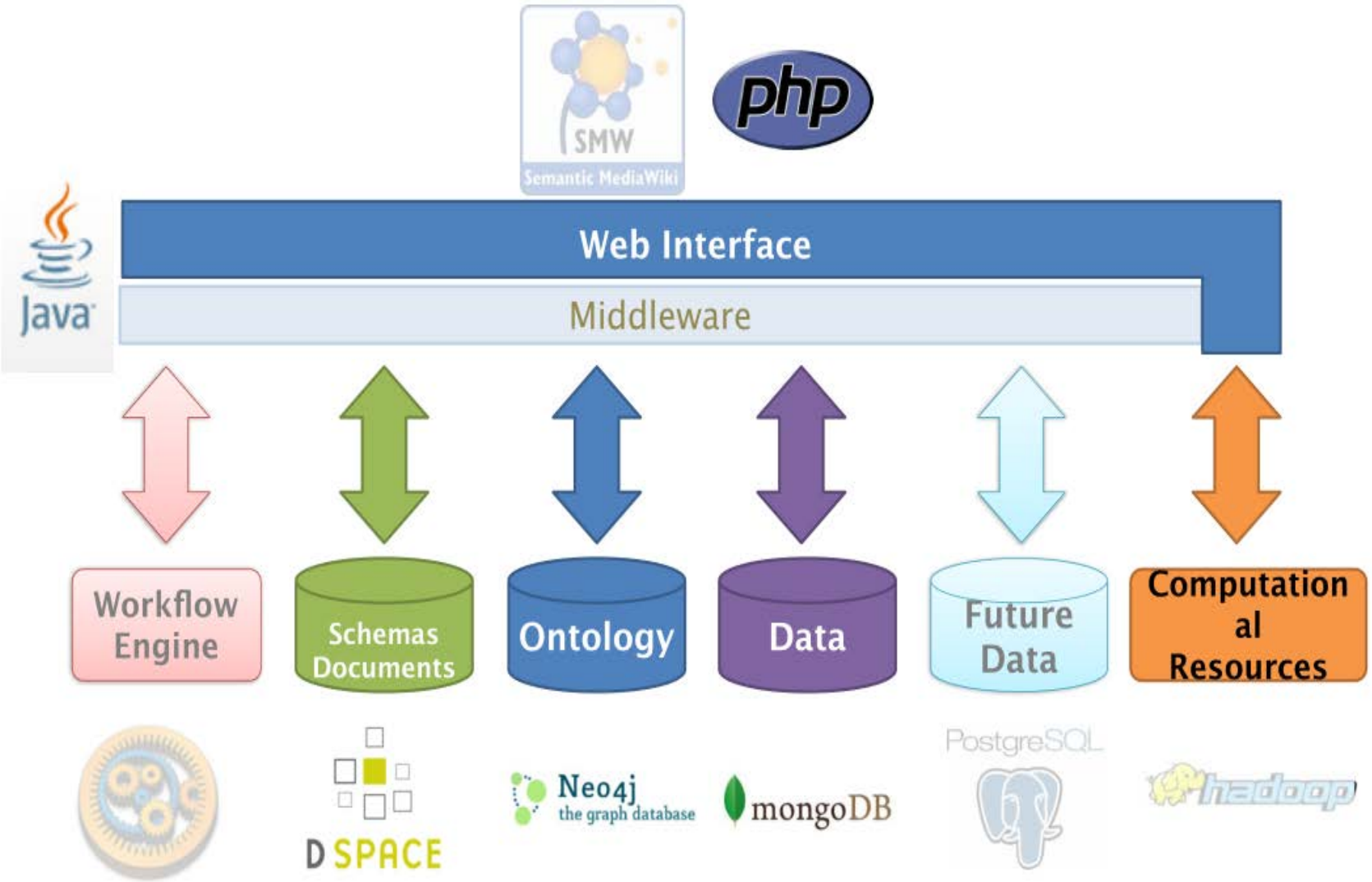




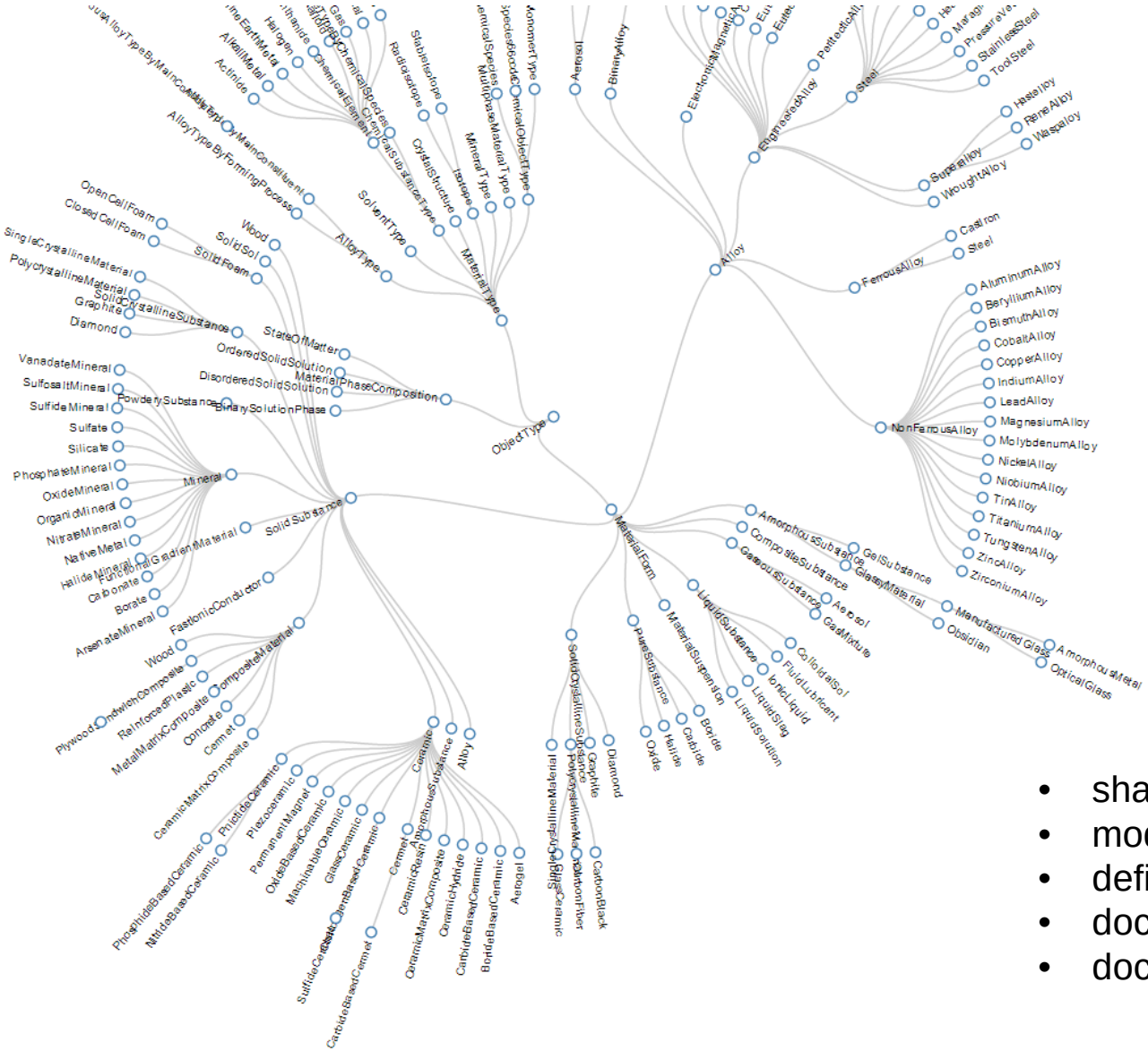
Customers and Partners

U.S. DEPARTMENT OF ENERGY – Energy Efficiency & Renewable Energy
QuesTek Innovations LLC
Oak Ridge National Laboratory
accelrys
GM
Carpenter
Lawrence Livermore National Laboratory
CompuTherm
Thermo-Calc Software
Ford
Materials Design
Kent State University

Planned MGI Architecture



Prototype MGI Ontology



Broad concepts covered in materials data files (data have many types)

- Objects, Materials, and Events
- Physical Properties
- Documents
- Data Objects & Types
- People & Organizations
- Software
- Relations among these

<u>Ontology</u>

- shared vocabulary and taxonomy
- models a domain
- defines objects and concepts
- documents object properties
- documents object relationships

MGI: Just a Special Case of Data Sharing

OSTP "Public Access" Memo
Feb 22, 2013

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles

The Administration is committed to ensuring that, to
constraints possible and consistent with law and the
federally funded scientific research are made availa
the scientific community. Such results include peer-

Scientific research supported by the Federal Govern
drive our economy. The results of that research beco
for progress in areas such as health, energy, the envi

Access to digital data sets resulting from federally f
resources and efforts on understanding and exploitin
data underpins the forecasting industry, and making
spawned many biotechnology innovations. In additi
publications and scientific data in digital formats wi
services related to curation, preservation, analysis, a
publications and data for re-use through preservatio
the impact and accountability of the Federal research
scientific breakthroughs and innovation, promote en
growth and job creation.

OMB "Open Data" Memo
May 9, 2013

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

officer

ficer

ce of Information and Regulatory Affairs

ging Information as an Asset

ource and a strategic asset to the Federal Government, its
at the Federal Government is taking full advantage of its
and agencies (hereafter referred to as "agencies") must
s life cycle to promote openness and interoperability, and
Managing government information as an asset will increase
e services, support mission needs, safeguard personal
luable government information.

ible, discoverable, and usable by the public can help fuel

Executive Order
May 9, 2013

For Immediate Release May 09, 2013

Executive Order -- Making Open and Machine Readable the New Default for Government Information

EXECUTIVE ORDER

- - - - - - -

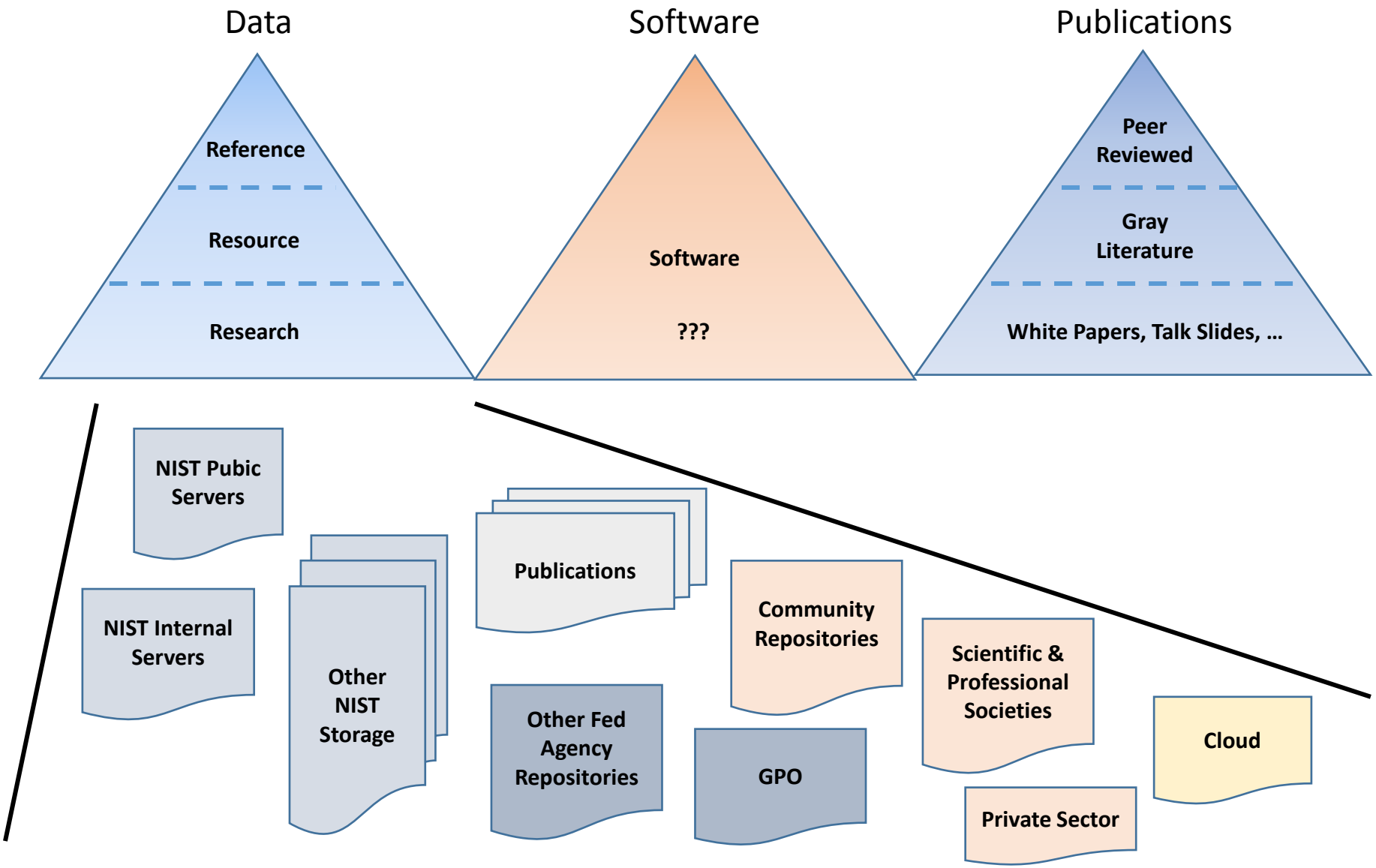
MAKING OPEN AND MACHINE READABLE THE NEW DEFAULT
FOR GOVERNMENT INFORMATION

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. General Principles. Openness in government strengthens our democracy, promotes the delivery of efficient and effective services to the public, and contributes to economic growth. As one vital benefit of open government, making information resources easy to find, accessible, and usable can fuel entrepreneurship, innovation, and scientific discovery that improves Americans' lives and contributes significantly to job creation.

Decades ago, the U.S. Government made both weather data and the Global Positioning System freely available. Since that time, American entrepreneurs and innovators have utilized these resources to create navigation systems, weather newscasts and warning systems, location-based applications, precision farming tools, and much more, improving Americans' lives in countless ways and leading to economic growth and job creation. In recent years, thousands of Government data resources across fields such as health and medicine, education, energy, public safety, global development, and finance have been posted in machine-readable form for free public use on Data.gov. Entrepreneurs and innovators have continued to develop a vast range of useful new products and businesses using these public information resources, creating good jobs in the process.

Scope of the Problem



Five Pilot Datasets: Standard Reference Data©

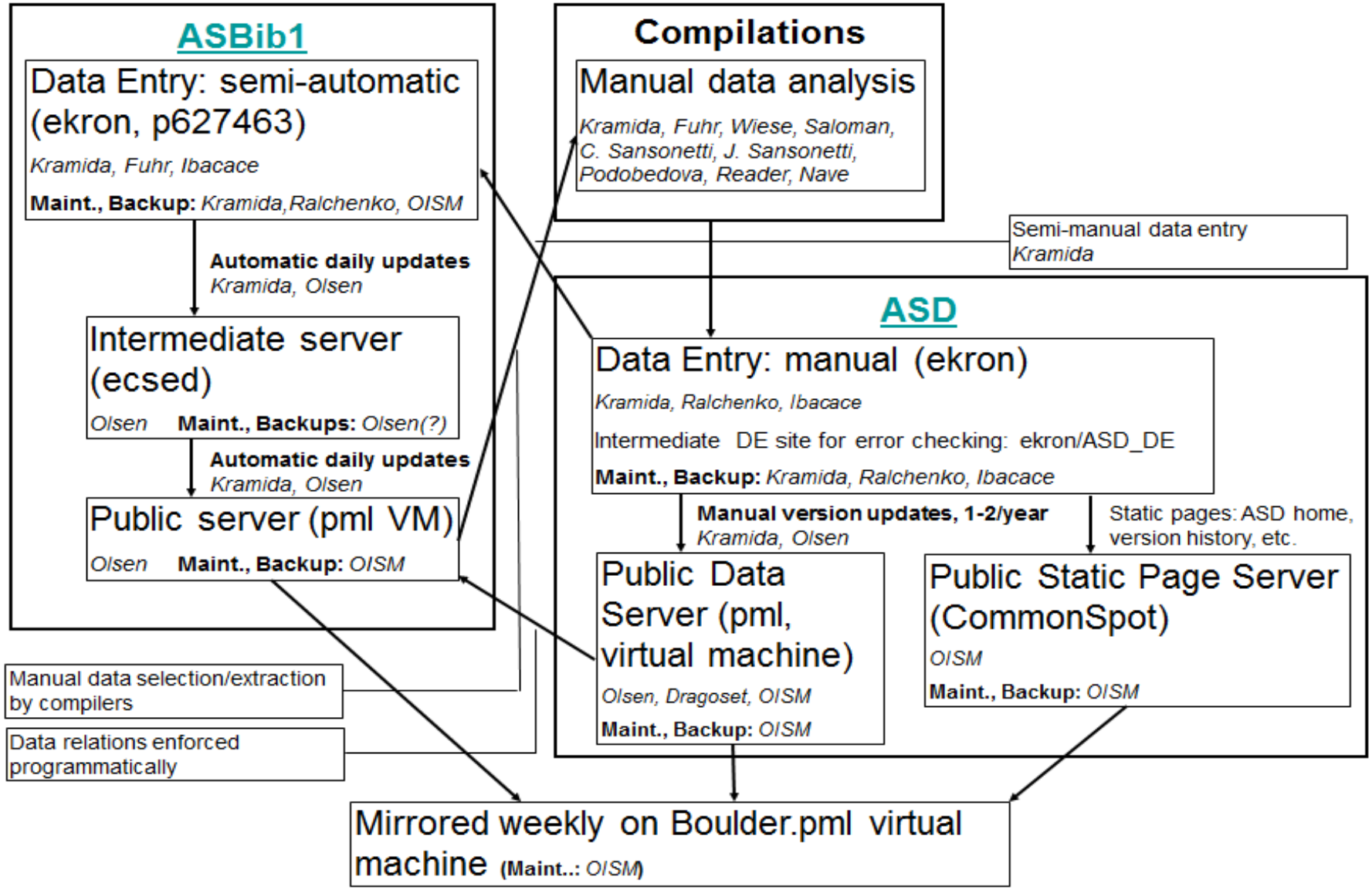
Database Name	SRD#	Code name	Authors	Contact person
Fundamental Physical Constants	121	FC	P.J. Mohr; B.N. Taylor; D.B. Newell	Peter J. Mohr
Atomis Spectra Database	78	ASD	A. Kramida; Yu. Ralchenko; J. Reader; NIST ASD Team	Alexander Kramida
NIST Chemical Kinetics Database	17		http://kinetics.nist.gov/kinetics/credits.jsp	Tom Allison
The NIST Chemistry WebBook	69		http://webbook.nist.gov/chemistry/	Peter Linstrom
Chem-Blast Gateway for PDB Ligands	155		T.N. Bhat	T.N. Bhat

Common Core ⟷ DataCite crosswalk spreadsheet

Two More Pilot Datasets: MGI Data

Database Name	Repository	URL	Contact person
NIST Calphad	DSpace	http://h058085.nist.gov:8080/xmlui/handle/123456789/1	Carelyn Campbell
Interatomic Potentials	Web	http://www.ctcms.nist.gov/potentials	Chandler Becker

Example 1: Atomic Spectral Database



ASD (Updated 1-2/year)

MySQL tables: 9
Data rows: 329046
(100% data)

Relations in MySQL & some programmatic;
Relations to ASBib1: programmatic

Web pages:
4 .cfm in CommonSpot
7 .html and 1 .css
on pml VM

Dump size: 75 Mb

Codes:
- Client-side: 5 .js (**18 Kb**) + javascript embedded in html and .pl + 1 Java code, **58 Kb**
- Server-side: 18 .pl and .cgi codes, **352 Kb**

ASBib1 (Updated daily)

MySQL tables: 29
Data rows: 540426
(89% metadata)

Relations in MySQL & programmatic.

Web pages:
2 .cfm in CommonSpot
1 .html **on pml VM**

Dump size: 100 Mb

Codes:
- Client-side: Javascript embedded in .cgi
  Server-side: 42 .pl and .cgi codes, **350 Kb**

Coding languages:	SQL, Perl, Java, Javascript, HTML, CSS, Latex
Database engine:	MySQL/ISAM
Web server:	Apache

Fundamental constants database

http://physics.nist.gov/cuu/Constants/index.html

Tables: **2929**
(tex, html, pdf, txt files, 14 Mb)

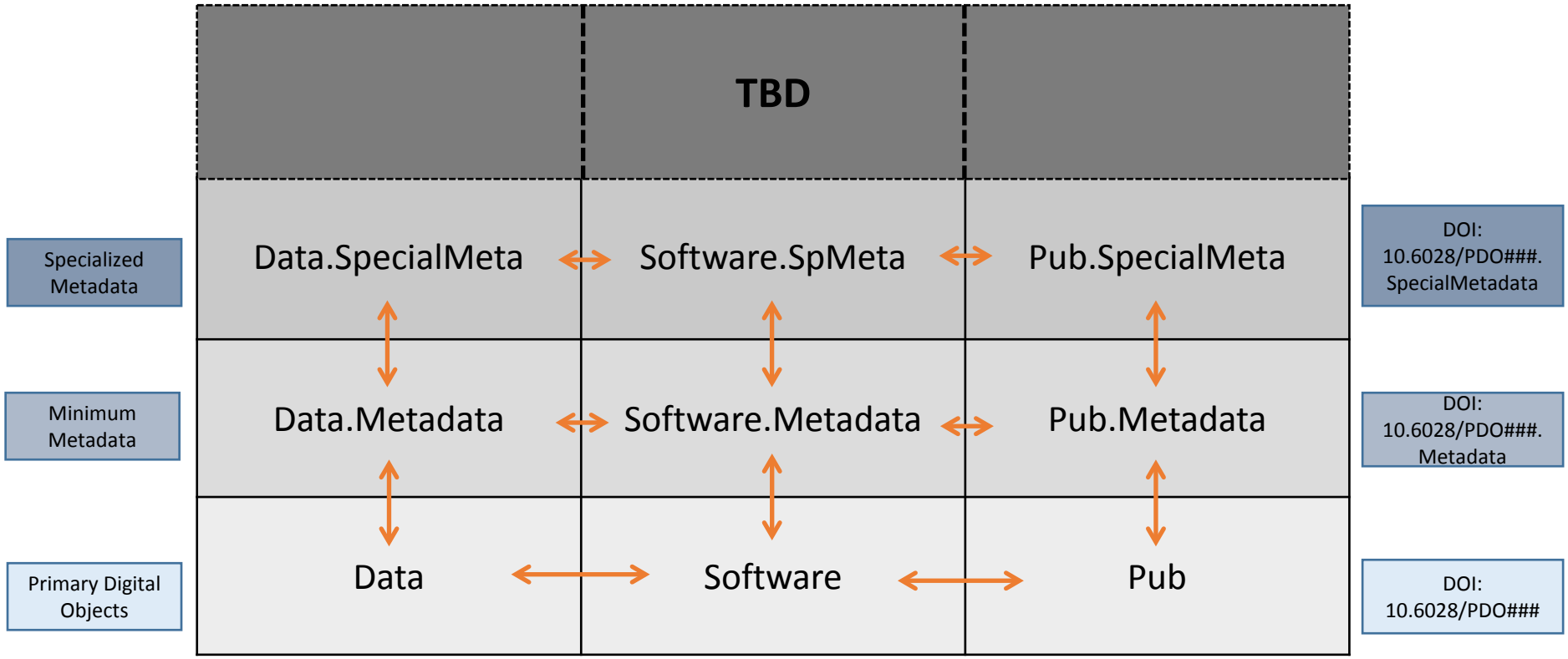
Web pages: **987**
(.html, .pdf, .txt)
28747 Mb

Codes (all server-side):
189 (2.2 Mb)
Perl, ksh, sed, awk, C, Fortran, css, html, tex

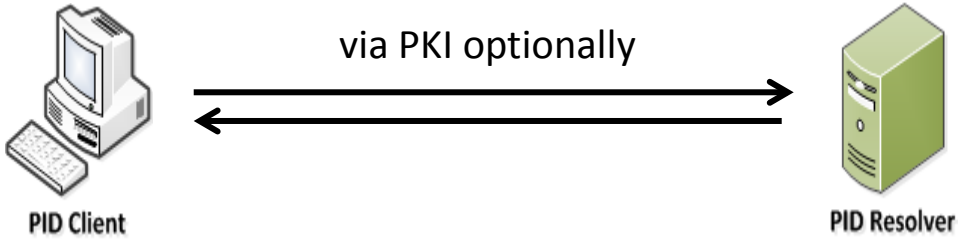
- Data dispersed throughout tables, web pages, and codes
- Executable codes dispersed throughout data and HTML files
- New versions released every ~ 4 years
- Old versions retained and re-linked to new versions via html links

Resides on PML VM, mirrored on a VM in Boulder

NIST Data Sharing Plan: Logical Model



Persistent Identifiers



10.12345/456		
1	PK	publickey
2	IP	rights data
3	Publisher	NIST
4	GUID	a8-0c-22-7f-c1-00
5	URL	http://pubmed.nih..
6	HDL	10.12345/9934
…	…	…

In this example the Handle System maps a persistent identifier to a set of metadata.

Query Type 1: Give me all data inside PID 10.12345/456
Query Type 2: Give me all data of type URL inside PID 10.12345/456

Persistent Identifiers (PIDs)

10.12345/456		
1	PK	publickey
2	IP	rights data
3	Publisher	NIST
4	GUID	a8-0c-22-7f-c1-00
5	URL	http://pubmed.nih..
6	HDL	10.12345/9934
…	…	…

To Do:
recommend what types of PID are needed (PID Information Types)

10.12345/9934		
1	PK	publickey
2	field1	xxx
3	field2	yyy
…	…	…

Common Access Platform

Johns Hopkins Sheridan Libraries

Corporation for National Research Initiatives

RDA
RESEARCH DATA ALLIANCE

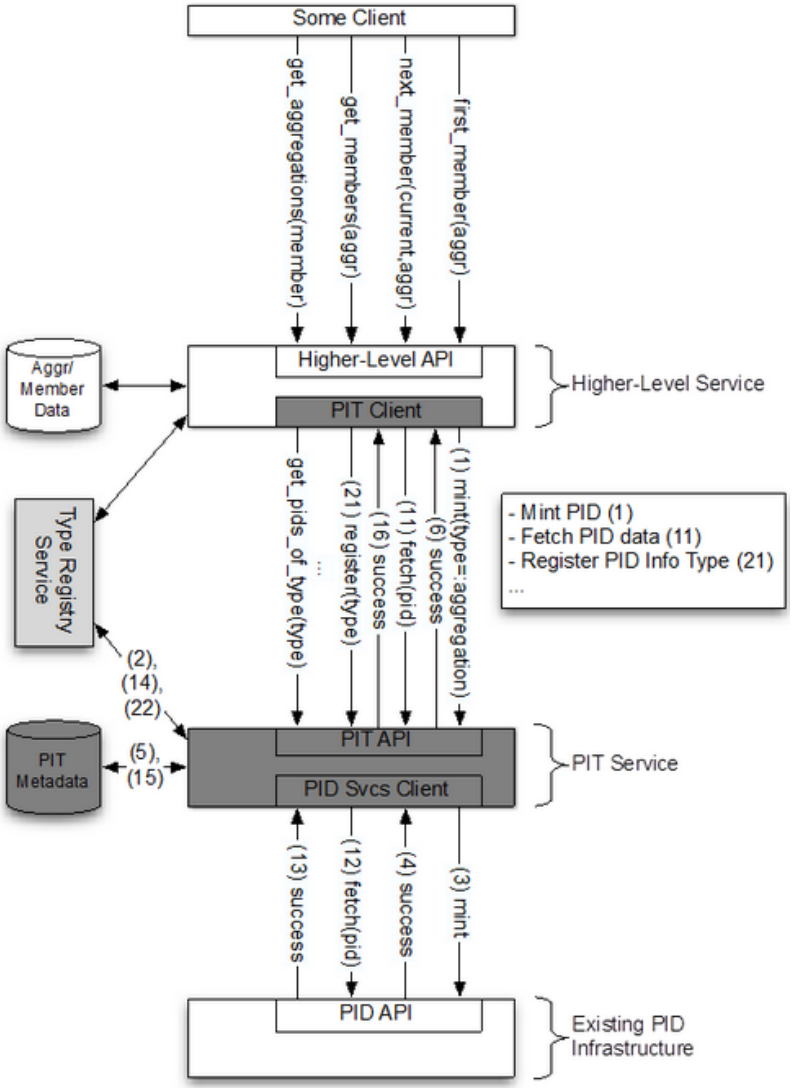
Research Data Sharing without barriers

PID Information Types WG

Status: Recognised & Endorsed
This Working Group focuses on cross-community concerns regarding Persistent Identifiers which are the core of proper data management and access.

Data Type Registries WG

Status: Recognised & Endorsed
The Data Type Registries Working Group will compile a set of use cases for data type use and management. It will identify and distinguish among existing 'type registry' efforts and their potential interaction with this group and formulate a data model and expression for types. It will also design a functional specification for type registries and propose a federation strategy among multiple type registries at both technical and organizational levels.



Common Access Platform

Johns Hopkins Sheridan Libraries

Corporation for National Research Initiatives

RDA
RESEARCH DATA ALLIANCE

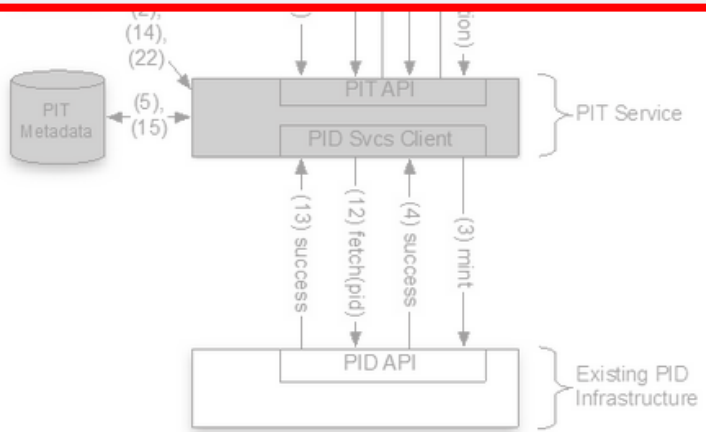
Research Data Sharing without barriers

Open Research Data ... breaking down barriers - RDA Second Plenary Meeting

16-18 September 2013 - National Academy of Sciences, Washington DC, US. Data sharing offers important benefits for scientific progress and advancement of knowledge. However, several limitations and barriers in the general adoption of data sharing are still in place.

Metadata Standards WG

Status: Submitted for Council review
The RDA Metadata Working Group identifies and endorses metadata solutions for addressing data management challenges. Metadata is crucial for the discovery, access, preservation, exchange, manipulation, and use/re-use of scientific data.



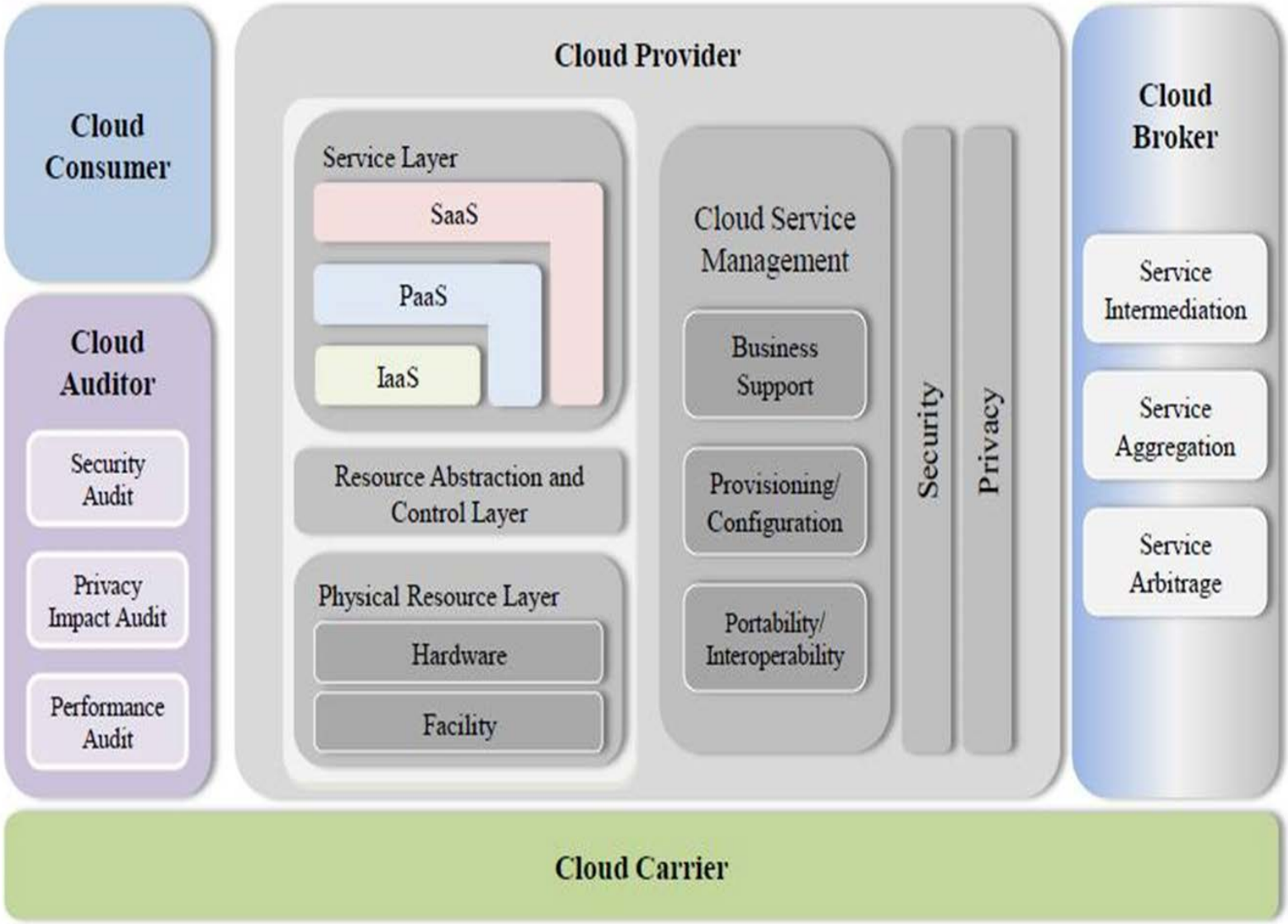
Use IETF as a model

- Challenges: interdisciplinary, interagency, lot of players bigger than NIST
- traditional Standards Development Organization (SDO) process
  - clean, elegant standards written by committee
  - often a near-optimal solution
  - issued from an ivory tower, promptly ignored by community
- IETF alternative process:
  - experience-based
  - documents and solidifies what has already been proven to work
  - testing and prototyping are central
  - de-emphasizes what *should* be done, “because we say so”
  - summed up in David Clark’s 1992 quote:

> “We reject kings, presidents, and voting.
>
> We believe in rough consensus and running code.”

David Clark
Internet Chief Protocol Architect
1981-1989

Why NIST ?


Cloud Consumer
Cloud Auditor
Security Audit
Privacy Impact Audit
Performance Audit
Cloud Provider
Service Layer
SaaS
PaaS
IaaS
Resource Abstraction and Control Layer
Physical Resource Layer
Hardware
Facility
Cloud Service Management
Business Support
Provisioning/ Configuration
Portability/ Interoperability
Security
Privacy
Cloud Broker
Service Intermediation
Service Aggregation
Service Arbitrage
Cloud Carrier

Acknowledgments

- NIST Scientific Data Committee
  - esp. Chris Greer, Wo Chang
- NIST Data Access Working Group
  - Peter Linstrom, Sasha Kramida, Andrea Medina-Smith, Kellie Beall, Dan Samarov, Bill Turner, Kirk Dohne, Adam Morey, Carolyn Rowland, Jonathan Hardis
- Larry Lannom, Giridhar Manepalli, Tim DiLauro