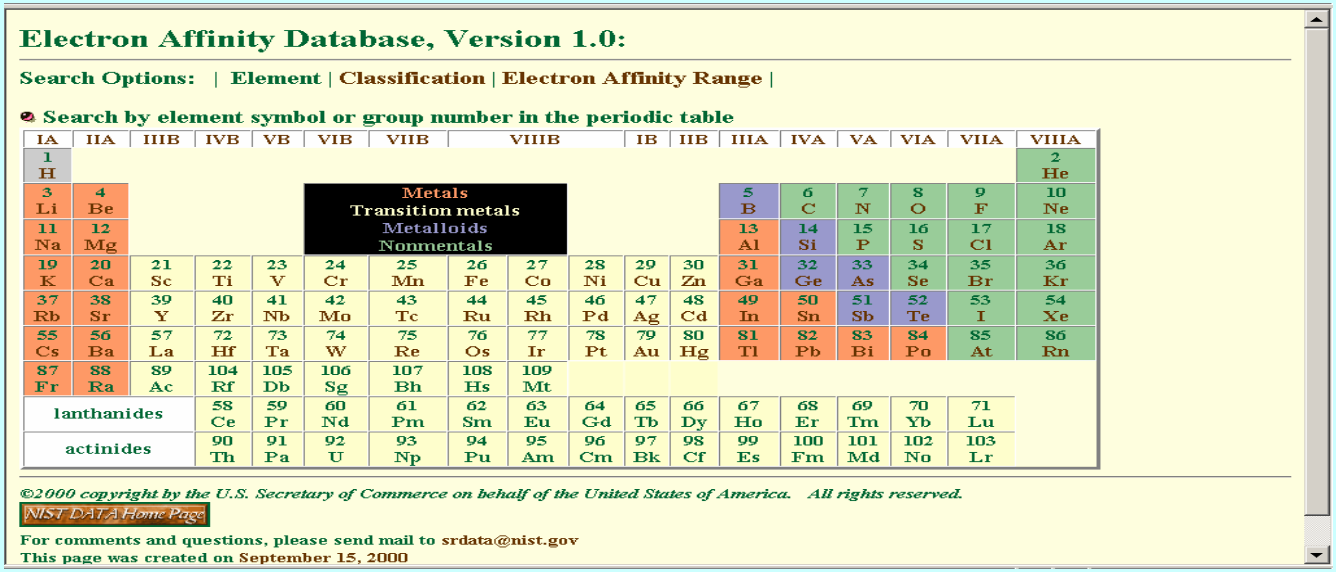


Preserving Scientific Data: New Methods for Scientific Discovery

John Rumble

National Institute of Standards and Technology

NIST
National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce

CODATA 2002

Data

“When you can measure what you are speaking about, and express it in numbers, you know something about it;

Lord Kelvin

Data

"When you can measure what you are speaking about, and express it in numbers, you know something about it;

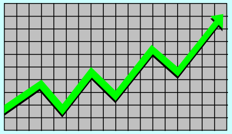
"But when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you scarcely in your thoughts advanced to the state of science."

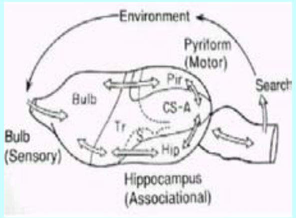
Lord Kelvin

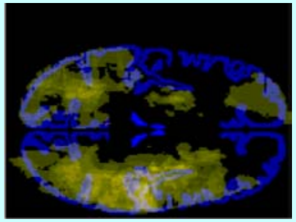
Types of Data

- Numbers
- Simple text
- Complex text
- Equations
- Graphs
- Diagrams
- Pictures
- Software
- Rules

- 1, 2, 3…
- ABCs
- Greek, scripts, symbol
- $E=mc^2$







Characteristics of Scientific Databases

Data

- From many publications or observations
- Full range of independent variables
- Large number of measurements
  - ➔ similar and varied
- Numbers of substances or systems
- Large amount of metadata

Text

- One or small number of studies
- Limited range of variables
- Small number of measurements
- Small number of substances or systems
- Small amount of metadata

Data Preservation and Scientific Discovery

- Data communicate measurement and calculation results
- Preserved data collections form the foundation of scientific discovery
- *Scientific discovery explains the observable world*

Data Preservation and Scientific Discovery

- Historical trends in data preservation and discovery
  - Accuracy
  - Comprehensiveness
  - Explanation of essence
  - Explanation of the complex
  - *Automated discovery - The future*

Accuracy

Newgrange – Ireland

- 6000 years old
- Aligned to the rising sun in the winter solstice
- Depended on careful observational data on the rising sun

Accuracy Improving

Stonehenge

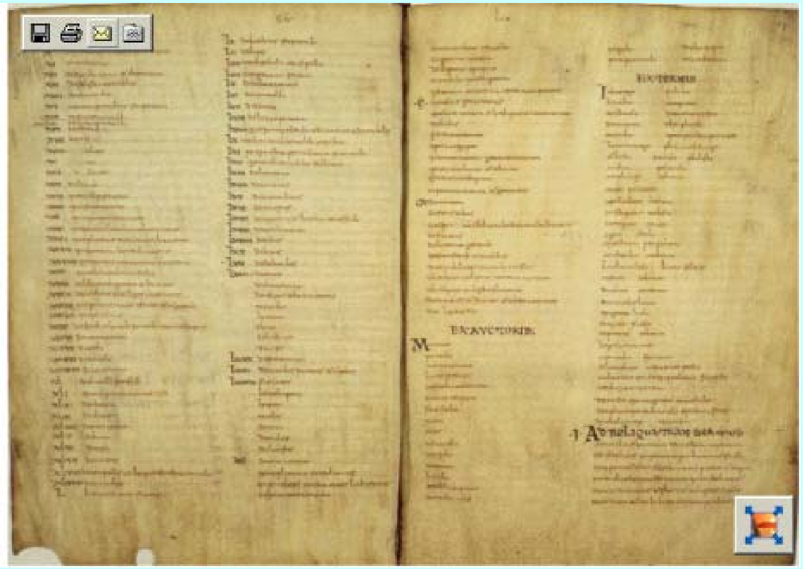
- 5000 years old
- Over 100 stones
- Complicated stone alignments
- Marks position of the moon and major stars as well as the sun
- Reproducibility of several observations

Comprehensive

Galen

- Greek physician
- Experimental physiologist
- Arabic copy from 800 AD
- Pictorial, descriptive, function describing
- Representative of botanical and animal catalogs

Comprehensive Improving



Pliny the Elder

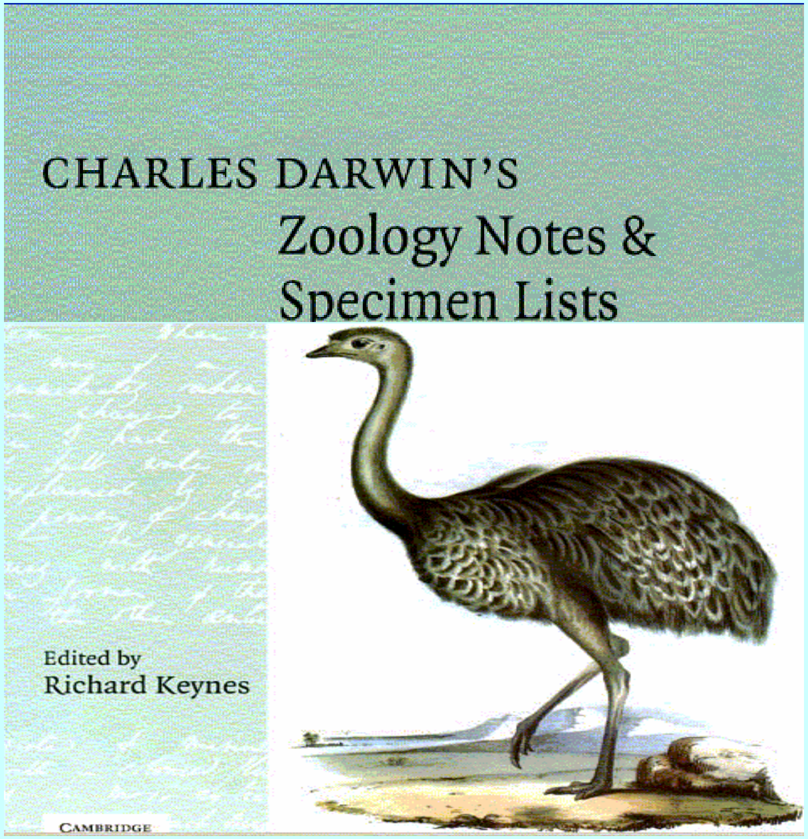
- Roman scholar
- *Natural History* (77 AD)
- One of earliest known encyclopedias of the natural world
- Systemization of data

Extraction of Essence

Tycho Brahae

- Late 16th Century
- Danish Astronomer
- Made precise measurements that led to Kepler's theories
- Led to discovery of simple relationships

Explanation of the Complex



Charles Darwin

- Combined with others in geology, zoology and botany
- A wide variety of facts and phenomena recorded
- Theory of Evolution had to explain all these observations and measurements

Prediction from Data

Electron Affinity Database, Version 1.0:

Search Options: | Element | Classification | Electron Affinity Range |

Search by element symbol or group number in the periodic table

Metals
Transition metals
Metalloids
Nonmentals

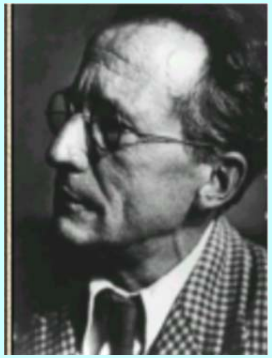
IA	IIA	IIIB	IVB	VB	VIB	VIIB	VIIIB			IB	IIB	IIIA	IVA	VA	VIA	VIIA	VIIIA
1 H																	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	57 La	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	89 Ac	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt									
lanthanides			58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu	
actinides			90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr	

NIST DATA Home Page

For comments and questions, please send mail to srdata@nist.gov
This page was created on September 15, 2000

Prediction from Data

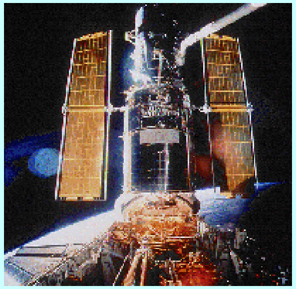
- Notes on the Spectral Lines of Hydrogen: Johann Jacob Balmer *Annalen der Physik und Chemie* 25 80-5 (1885)
- I gradually arrived at a formula which, at least for these four lines, expresses a law by which their wavelengths can be represented by striking precision…From the formula, we obtained for a fifth hydrogen line 3936.65x10-7 mm.
- The development of quantum mechanics
- Bohr and Schrödinger



Data and Experimentation Today

Today we have exciting new capability to observe nature better than ever before

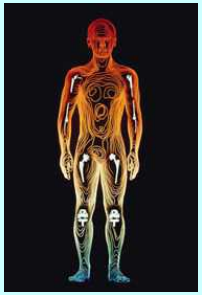
- ➔ Atomic force microscopes
- ➔ Hubble Space Telescope
- ➔ Micro-electronics and lasers
- ➔ High power computers to analyze data
- ➔ Biomacromolecule sequencing instruments



Generates large amounts of quality data

Data and Computation Today

We now also have the ability to create a Virtual World



- Models and simulations of complex systems
- Techniques to do advanced mathematics
- Computers to execute immense calculations
- Visualization tools to examine our virtual world

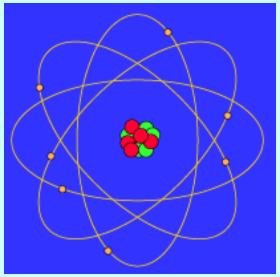
Requires and generates large amounts of quality data

Data and The Information Revolution

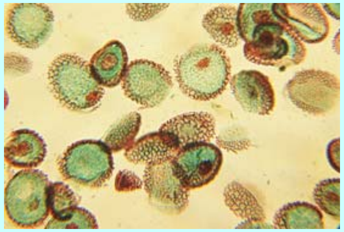
- Computer at every desk
- The Internet/WWW explosion
- Database tools on every computer
- Electronic publications
- Model and simulation-based R&D
- Virtual libraries
- Comprehensive databases

- ***Data at the very heart of the revolution***

21st Century Science



- From the fundamental to the complex
  - ➔ Determining the laws of nature for a few particles to understanding *real systems* - cells, the atmosphere, the Earth, ecology
- From reductionism to constructionism
  - ➔ Using our basic knowledge to make models and predict behavior of *real systems*



The Face of 21st Century Science

- Complex
- Multi-disciplinary
- Real systems
- Virtual as well as physical

Access to quality data becomes critical

Long term preservation of and access to data becomes more important than ever!

Major Point of Today's Talk

Scientific databases in the future will be even more important source for scientific discovery

- Preservation of data needed for
  - New insights
  - Scientific principles
  - New knowledge
  - Understanding complex systems

And the discovery will be computer-aided, if not done by computers alone

Data Preservation in the Future

Yesterday

- Collections managed by a small number of people
- Collections readable by one scientist
- Collections interpretable by one person
- Discoveries made by thinking, with analysis by one person

Future

- Collections managed by groups
- Collections not readable by any individual
- Collections interpretable only with aid of software
- Discoveries made by computers, with verification by people

Discovering Science in Preserved Data

Real systems are very complex

- Large number of objects
- Large number of independent variables
- Collective behavior difficult to find
- Abstraction of important features
- Existence of unifying theory or concept
- Multiple views

Discovering Science in Preserved Data

- Too much data for any one *person* to understand

  How long does it take to look over a terabyte of data?

Discovering Science in Preserved Data

Real systems are very complex

- Large number of objects - mole, species, stars, geographic points

- *How much data is needed to come up with an idea?*
- *Does quality count?*

Discovering Science in Preserved Data

Real systems are very complex

- Large number of independent variables

- *How do we use metadata to describe what we preserve?*
- *How do they change over time and context?*
- *If we must aggregate different data sets (e.g., over the Web) to do discovery, how do we know data are comparable?*

Discovering Science in Preserved Data

Real systems are very complex

- Collective behavior difficult to find

- *How do we recognize real phenomena from artifacts?*
- *What kind of data visualization and exploitation (discovery) tools will exist in 20 years?*
- *Weather prediction for the next day!*

Discovering Science in Preserved Data

Real systems are very complex

- Abstraction of important features
- *How can we find what is important when we have too much data?*
- *Cholesterol linkage to heart disease was found by computer-aided correlation.*

Discovering Science in Preserved Data

Real systems are very complex

- Existence of unifying theory or concept
- *Could we derive quantum mechanics from a complete database of atomic and molecular spectra?*
- *What features does quantum mechanics have beyond these data?*

Discovering Science in Preserved Data

Real systems are very complex

- **Multiple views**

- ***Quantum theory, matrix mechanics, Maxwell's theory; quantum electrodynamics***

- ***Are all views of nature equally discoverable?***

Important New Data Collections

- International Virtual Observatory
- Structural Genomics
- Proteomics
- Climate change

- Historic geologic
- Chemistry on demand

- Biodiversity
- Brain scans

- All observation for every point in the sky
- For living things!
- For all living things
- Water, earth, atmosphere and all they contain

- Lots of years, lots of rocks
- 60 elements, 5 at a time, different ratios, ???
- 5M species? or 10M? or 50M
- Just think, forever

Challenges of the Data Era

The technology to handle the overwhelming volume of data from new measurement techniques

- What to capture when sensors generate too much too fast?
- How to store, represent, manipulate and display too voluminous data?
- How to find out which data are important?

Challenges of the Data Era

Making accurate virtual measurements on virtual systems

- What is uncertainty in a calculation?
- How do you establish traceability for a calculation?
- What computational results should be stored, and how can those data be handled?

Challenges of the Data Era

Evaluating data quality

- How can large amounts of data be evaluated? In real time? As new data are published?
- How can large data sets be integrated together correctly?
- How do you determine the quality of a calculation?
- What does quality mean in a terabyte of data?

Challenges of the Data Era

Making exploitation of large data sets possible

- What standards are needed for making data sets work together?
- How can you verify discovery from data sets?
- How can you make control decisions when you have too much data?

Challenges of the Data Era

How do we maintain full and open access to the large number of databases required for making new scientific discoveries

- What policies are needed for full and open access?
- How can discoverers profit from their automated discoveries?
- How do you get the information industry to understand the new paradigm for discovery?

Some Final Thoughts

Scientific databases in the future will be even more important source for scientific discovery

- Preservation of data needed for
  - New insights
  - Scientific principles
  - New knowledge
  - Understanding complex systems

Will computers discover and people just verify?

Some Final Thoughts

- *Let's take advantage of CODATA's expertise, neutrality and openness to support scientific and technological advances in the future*