Seamless Astronomy
Alyssa A. Goodman
Harvard-Smithsonian Center for Astrophysics

with
Alberto Accomazzi, Douglas Burke, Gus Muench, Michael Kurtz & Alberto Pepe (Harvard-Smithsonian CfA); Eli Bressert (U. Exeter); Tim Clark (Massachusetts General Hospital/Harvard Medical School); Chris Borgman (UCLA); Jonathan Fay & Curtis Wong (Microsoft Research)
“Standard” Practices in Bioinformatics?

PubMed Central

GenBank Overview

Growth of GenBank

What is GenBank?

GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research, 2008 Jan 36(Database issue):D25-30). There are approximately 106,533, 56,756 bases in 108,431,692 sequence records in the traditional GenBank divisions and 148,185,177,763 bases in 48,443,067 sequence records in the WGS division as of August 2009.

The complete release notes for the current version of GenBank are available on the NCBI FTP site. A new release is made every two months. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

An example of a GenBank record may be viewed for a Saccharomyces cerevisiae gene (Swine Flu). The Centers for Disease Control and Prevention and other health officials are actively tracking the recent emergence of human cases of swine influenza A (H1N1) virus infection. Influenza A virus sequences from patients affected by this strain are being submitted to Genbank and can be accessed through the NCBI PubMed database.
“Standard” Practices in Bioinformatics?

NCBI Overview

GenBank Overview

■ What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research, 2008 Jan 36(Database issue):D25-30). There are approximately 106,533,156,756 bases in 108,431,692 sequence records in the traditional GenBank divisions and 148,185,117,763 bases in 48,443,967 sequence records in the WGS division as of August 2009.

The complete release notes for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

An example of a GenBank record may be viewed for a Saccharomyces cerevisiae gene.

In The News: 2009 H1N1 Flu Virus (Swine Flu)

The Centers for Disease Control and Prevention and other health officials are actively tracking the recent emergence of human cases of swine influenza A (H1N1) virus infection. Influenza A virus sequences from patients affected by this strain are being submitted to GenBank and can be accessed through the NCBI Flu Resource.
Bioenergetic Profile Experiment using C2C12 Myoblast Cells
David G. Nicholls1, Victor M. Darley-Usmar2, Min Wu3, Per Bo Jansen1, George W. Rogers3, David A. Ferrick3
1Buck Institute for Age Research, Novato, CA, 2Department of Pathology, Center for Free Radical Biology, University of Alabama at Birmingham - UAB, 3Seashore Bioscience, North Billerica, MA

A description of a method for profiling mitochondrial function in cells is provided. The mitochondrial profile generated provides four parameters of mitochondrial function that can be measured in one experiment: basal respiration rate, ATP-linked respiration, proton leak, and reserve capacity.

Non-invasive Imaging of Leukocyte Homing and Migration in vivo
Baomei Wang1, Bernd H. Zinselmeyer2,1, Jeremiah R. McDole1, Peggy A. Gieselmann1, Mark J. Miller3
1Department of Pathology and Immunology, Washington University in St. Louis, 2National Institute of Neurological Disorders and Stroke, NINDS, NIH - National Institute of Health

Here, we describe a non-invasive two-photon (2P) microscopy approach to study leukocyte homing in the mouse footpad. We discuss the technical aspects of our tissue imaging preparation and walk the reader through a typical experiment from initial set up to execution and data collection.
Realm of Seamless Astronomy
3500 years of Observing

Stonehenge, 1500 BC

Ptolemy in Alexandria, 100 AD

Observatory Tower, Lincolnshire, UK, c. 1300

Galileo, 1600

The “Scientific Revolution”

Reber’s Radio Telescope, 1937

NASA/Explorer 7 (Space-based Observing) 1959

“The Internet”

Long-distance remote-control/“robotic” telescopes 1990s

“Virtual Observatories” 21st century
What can today’s Astronomer’s “Research” look like?

Research

In my Astronomy research, I am primarily interested in how the gas in galaxies constantly re-arranges itself over huge time spans to constantly form new stars. I have also had a long-standing interest in data visualization, and in improving the use of computers in all aspects of scientific research. I teach a course at Harvard called “The Art of Numbers,” and I am very involved in the WorldWide Telescope Project, which brings astronomical data to everyone through an interface that demonstrates data delivery for the 21st Century of "e-Science."

http://www.cfa.harvard.edu/~agoodman/
**World Wide Telescope:** a UIS from Microsoft Research

[UIS=Universe Information System]

Created by Curtis Wong and Jonathan Fay at MSR; AG is “Academic Partner” on the WWT Project

Seamless *Data/Literature Connections* (e.g. ADS)

“*Modular Craftsmanship*” (e.g. flickr)

Collections, Communities & *Guided Tours*

The World Wide Telescope
an Archetype for Online-Science

Jim Gray (Microsoft)
Alex Szalay (Johns Hopkins University)
Microsoft Academic Days in Silicon Valley

http://research.microsoft.com/~gray/talks
The (US) Backstory


Science News

$10 Million NVO Grant

ScienceDaily (Oct 25, 2010) - NVO (the Universe at your fingertips) is the world's largest research institute, starting an ambitious new universe online.

See Also:

NVO senior personnel:
Charles Alcock, University of Pennsylvania Kirk Burke, Astronomical Data Center/Raytheon Tim Cornwell, NSF National Radio Astronomy Observatory Optical Astronomy Observatory Giuseppe Fabbiano, Smithsonian Astrophysical Observatory Alyssa Goodman, Harvard University Jim Gray, Microsoft Robert Hanisch, Space Telescope Science Institute George Helou, NASA Infrared Processing and Analysis Center Stephen Kent, Fermilab Carl Kesselman, University of Southern California Carl Kesselman, University of Illinois, Urbana-Champaign

Management and Operation of the Virtual Astronomical Observatory

CONTACTS

Name Email
Nigel Sharp nsharp@nsf.gov
Eileen D. Friel efriel@nsf.gov

PROGRAM GUIDELINES

Solicitation 08-537

Please be advised that the NSF Proposal & Award Policies & Procedures Guide (PAPPG) includes revised guidelines to implement the mentoring provisions of the America COMPETES Act (ACA) (Pub. L. No. 110-69, Aug. 9, 2007) as specified in the ACA, each proposal that requests funding to support postdoctoral researchers must include a description of the mentoring that will be provided for such individuals. Proposals that do not comply with this requirement will be returned without review (see the PAPPG Guide Proposal Guide Chapter II for further information about the implementation of this new requirement).

NVO
...the Universe at your fingertips

VAO
VIRTUAL ASTRONOMICAL OBSERVATORY

and meanwhile...

Welcome to AstroGrid
AstroGrid is the doorway to the Virtual Observatory (VO). We provide a suite of tools to enable astronomers to explore and bookmark resources from around the world, find data in VOSpace, query databases, plot and manipulate tables, cross-match catalogues, and automate sequences of tasks. Tools from other Euro VO projects inter-operate with AstroGrid.

Virtual Observatory Software for Astronomers
HOME INSTALL HELP SUPPORT

Welcome to Aladin
Aladin is an interactive software sky atlas allowing the user to visualize digitized astronomical images, superimpose entries from astronomical catalogues or databases, and interactively access related data and information from the Simbad database. The VO/VISS and other archives for all known sources in the field can be displayed.

Created in 1998, Aladin has become a widely-used VO portal capable of addressing challenges such as locating data of interest, accessing and exploring distributed datasets, visualizing multi-wavelength data. Compliance with existing or emerging VO standards, interconnection with other visualization or analysis tools, ability to easily compare heterogeneous data are key features allowing Aladin to be a powerful data exploration and integration tool as well as a science explorer.

The Aladin Sky Atlas
Description
Aladin is an interactive software sky atlas allowing the user to visualize digitized astronomical images, superimpose entries from astronomical catalogues or databases, and interactively access related data and information from the Simbad database.

Copyright © 2009, De Re Studenten Vereniging, Universiteit Utrecht

The VO
What/where are/is “Data”?
What/where is literature?
What is a Virtual Observatory?
Seamless Astronomy

But, that was 2009...
Realm of “Seamless Astronomy”

Data

Advanced Search & InfoViz tools

literature

Literature

“Researcher”

Standalone Analysis Software

2010 Evermore Seamless Astronomy
This simple argument, first made at the 2009 WWT session at AAS, seems to be working:

“Astronomy research tools should work as seamlessly as travel research tools.”
“Astronomy research tools should work as seamlessly as travel research tools.”

When the concept of a "Virtual Observatory" (VO) was first discussed by future-looking astronomers in the mid-1990s, all thoughts were about distributed data and a common system to access it. But, information access on today's web primarily works in the reverse: distributed tools accessing common data centers. Capability and ease-of-use improvements to the web typically now come in the form of nesting, aggregating or connecting tools. Think kayak.com, iGoogle, or Bing Maps. In the "Seamless Astronomy" view to be discussed, today's "VO" should be thought of as the ever-improving set of data archives, tools, interconnections, and standards that strive to make astronomical research as "seamless" as travel research. The good news is that the cutting-edge of the astronomical research environment is moving rapidly in this seamless direction. The most savvy institutions are beginning to realize that the original VO model of data distributed on thousands of individual researchers' desktop hard drives is not a sustainable model, and that they need to offer data hosting, archiving, and stewardship services the way libraries offer such services for printed matter. Software tools are becoming much more interoperable thanks to protocols for message-passing such as "SAMP." And, the improved speed of web applications is to some extent removing platform-dependence as an obstacle to programmers and users alike. The bad news is that most astronomers are largely unaware of the tools that this new nirvana offers, and instead still conduct online research in the same way they did a decade ago. In this talk, I will focus in particular on how our recent work on connecting Microsoft's WorldWide Telescope program to other commonly-used astronomical research tools--most notably literature searching tools--has made the astronomical research environment more seamless. More generally, I will emphasize and demonstrate that an ever-increasing diversity of tools allow researchers to carry out a particular research task, so that the important research for the future lies in figuring out how to make the tools, their interconnections, and their connections to data and literature resources useful and well-known to the astronomical community.

(Abstract of the Evermore Seamless Astronomy presentation by A. Goodman at the Microsoft External Research Symposium, April 2010.)
Astronomers can see parallels...
Literature Handling: Diverse Apps, Common Data
What fraction of astronomy researchers know about these tools?
“writemypaper.org?”
“writemypaper.org?”
World Wide Telescope

NGC 7023

Classification: Reflection Nebula in Cepheus

RA: 21h01m36s
Dec: +68° 10' 11"
Dist: n/a
Magnitude: n/a
Alt: 30° 55' 38"
Az: 341° 36' 56"
Rise: Circumpolar
Set: Circumpolar
Transit: Circumpolar

Image Credits: Jack Newton

http://www.jacknewton.com/
<table>
<thead>
<tr>
<th>#</th>
<th>Bibcode</th>
<th>Authors</th>
<th>Score</th>
<th>Date</th>
<th>Title</th>
<th>List of Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2009ApJ...700.1609M</td>
<td>Myers, Philip C.</td>
<td>1.000</td>
<td>08/2009</td>
<td>Filamentary Structure of Star-forming Complexes</td>
<td>A Z E F L X</td>
</tr>
<tr>
<td>2</td>
<td>2009ApJ...700.1190D</td>
<td>Desai, Vandana; Soifer, B. T.; Dey, Arjun; LeFloc'h, Emeric; Armus, Lee; Brand, Kate; Brown, Michael J. L.; Brodin, Mark; Jannuzi, Buell T.; Houck, James R.; and 8 coauthors</td>
<td>1.000</td>
<td>08/2009</td>
<td>Strong Polycyclic Aromatic Hydrocarbon Emission from z ≈ 2 ULIRGs</td>
<td>A Z E F L X</td>
</tr>
<tr>
<td>3</td>
<td>2009MNRAS.396.1851N</td>
<td>Nutter, D.; Stamatellos, D.; Ward-Thompson, D.</td>
<td>1.000</td>
<td>07/2009</td>
<td>The initial conditions of isolated star formation - IX. Akari mapping of an externally heated pre-stellar core</td>
<td>A Z E F L X</td>
</tr>
</tbody>
</table>
Basic data:
NGC 7023 -- Open (galactic) Cluster

Other object types: CI+ (C,CI),(BD2003)) , Ope (OCISM) , HII (LBH) , V* (AAVSO) , IR (IRAS)

ICRS coord. (ep=2000): 21 01 36.9 +68 09 48 (-) [---] D --
FK5 coord. (ep=2000 eq=2000): 21 01 36.9 +68 09 48 (-) [---] D --
104.0616 +14.1926 (-) [---] D --

Fluxes (1):
B 7.20 [-] D --

Identifiers (11):
NGC 7023
C 2059+679
CI VDB 139
IRAS 20599+6755
IRAS P20599+6755
LBH 487
OCIS 50
AAVSO 2044+67

Plots and Images

References (371 between 1983 and 2009)
Simbad bibliographic survey began in 1950 for stars (at least bright stars) and in 1983 for all other objects (outside the solar system).
“writemypaper.org?”
And now we got to NGC 7023 by using the literature as a filter.
Seamless “Registration” through...

flickr + astrometry.net + WWT !?
Coming Soon

Historical Image Layer
Extracted from ALL ADS holdings (using astrometry.net)

Faceted Heat Map of Articles on the Sky

ADS-Seamless-astrometry.net collaboration
ADS-CDS-Seamless collaboration
Prototype of Articles on the Sky (April 2010)

Simbad bibliographical map (made by T. Boch CDS)

with thanks to CDS/Pierre Fernique
The future (is here)... data IN articles

Note: This work came from the “AstroMed” project am.iic.harvard.edu

data, CLUMPFIND typically finds features on a limited range of scales, above but close to the physical resolution of the data, and its results can be overly dependent on input parameters. By tuning CLUMPFIND’s two free parameters, the same molecular line data can be used to show either that the frequency distribution of clump mass is the same as the initial mass function of stars or that it follows the much shallower mass function associated with large-scale molecular clouds (Supplementary Fig. 1).

Four years before the advent of CLUMPFIND, “structure trees” were proposed as a way to characterize clouds’ hierarchical structure using 2D maps of column density. With this early 2D work as inspiration, we have developed a structure-identification algorithm that abstracts the hierarchical structure of a 3D (\(p_{\text{DR}}\)) data cube into an easily visualized representation called a “dendrogram”\(^{10}\). Although well developed in other data-intensive fields\(^{11,12}\), it is curious that the application of tree methodologies so far in astrophysics has been rare, and almost exclusively within the area of galaxy evolution, where “tree trunks” are being used with increasing frequency\(^{12}\).

Figure 3 and its legend explain the construction of dendrograms schematically. The dendrogram quantifies how and where local maxima of emission merge with each other, and its implementation is explained in Supplementary Methods. Critically, the dendrogram is determined almost entirely by the data itself, and it has negligible sensitivity to algorithm parameters. To make graphical presentation possible on paper and 2D screens, we “flatten” the dendrograms of 3D data (see Fig. 3 and its legend), by sorting their “branches” to not cross, which eliminates dimensional information on the z-axis while preserving all information about connectivity and hierarchy. Numbered “bellied ball” labels in the figures let the reader match features between a 2D map (Fig. 1), an interactive 3D map (Fig. 2a online) and a sorted dendrogram (Fig. 2c).

A dendrogram of a spectral-line data cube allows for the estimation of key physical properties associated with volumes bounded by isosurfaces, such as radius (\(R\)), velocity dispersion (\(\sigma\)), and luminosity (\(L\)). The volume can have any shape, and in other work\(^{13}\) we focus on the significance of the especially elongated features seen in L1448 (Fig. 2a). The luminosity is an approximate proxy for mass, such that \(M_\text{clump} \approx \frac{L_{\text{iso}} \times V}{\frac{1}{2} \times \text{gas mass}}\), where \(L_{\text{iso}} = 8.0 \times 10^{-12} \text{ erg K}^{-1} \text{ s}^{-1}\)’s (ref. 15, see Supplementary Methods and Supplementary Fig. 2). The derived values for size, mass and velocity dispersion can then be used to estimate the role of self-gravity at each point in the hierarchy, via calculation of an “observed” virial parameter, \(A_{\text{vir}} = \frac{V^2}{2GM} R\), where \(A_{\text{vir}} < 2\) (where gravitational energy is comparable to or larger than kinetic energy) correspond to regions of \(p_{\text{DR}}\) space where self-gravity is significant. As \(A_{\text{vir}}\) only represents the ratio of kinetic energy to gravitational energy at one point in time, and does not explicitly capture external over-pressure and/or magnetic fields\(^{16}\), its measured value should only be used as a guide to the longevity (boundedness) of any particular feature.

\[ \frac{1}{2}M \lambda \]
“Old Data” \[\text{astrometry.net/flickr/WWT}\] \[\text{3D PDF}\] \[\text{astrometry.net/flickr/WWT}\] "New Data"

“Your Data” \[\text{astrometry.net/flickr/WWT}\] \[\text{3D PDF}\] WWT as API "My Data"

"Old Data" WWT/ADS/SIMBAD/VAO WWT/ADS/SIMBAD/VAO WWT as API

"My Data"
Jim Gray & Alex Szalay had it right (in 2004)

All Scientific Data Online

- Many disciplines overlap and use data from other sciences
- Internet can unify all literature and data
- Go from literature to computation to data back to literature
- Information at your fingertips for everyone-everywhere
- Increase Scientific Information Velocity
- Huge increase in Science Productivity
How do we increase the fraction of astronomy researchers who know about these tools?

User Groups (CfA now has one)

+Suggestions?!
User Groups
(CfA now has one)
How do we increase the number of people who create and interlink new tools?

Kiva model: WWT Partners & “VAO Associates”
How do we organize such diverse tools, so as to make them interoperably useful?....

“SAMP” is a great technical start, but offers a very significant user interface challenge.
SAMP
Think about the “modules” needed to make this work...but do the details matter, to your research, if the system works seamlessly?
Seamless Astronomy

Mockup based on work of Eli Bressert, excerpted from NASA AISRP proposal by Goodman, Muench, Christian, Conti, Kurtz, Burke, Accomazzi, McGuinness, Hendler & Wong, 2008
Fiction (for now) but, coming very soon, e.g. ASCOT

Fact (right now)
“Faceted Browsing”

- create intelligent applications which can reason and infer with data
- publish resources as Linked Data, externally indexed
- easily aggregate metrics of interest to publishers, funding agencies
- let others build applications on this substrate using SPARQL queries

We will be carrying out these efforts as part of ADS Labs.

Bootstrapping in ADS Labs

ADS Labs is an effort to put out more forward thinking, somewhat unstable applications will be incubated in ADS Labs before being pushed out to ADS.

1. The results of queries on a bibliographic database will be made available in a data store. We will build a user interface on the above (see first image below).
2. We will switch to a semantic backend with a SPARQL interface.
3. Development on Ontologies (which this site details) continues and
4. Finally we'll combine the databases so as to have one large semantic store.

Examples of Applications

Here are examples of what such applications might look like:

Ongoing “ADS Labs” Work: Alberto Accomazzi, Rahul Davé, Alberto Pepe, Michael Kurtz, Gus Muench, AG

Thanks to ADS (NASA)/VAO (NASA+NSF)/MSFT funding.
Article Markup via ADS will be similar to working Life Sciences Tool (thanks IIC!)

"Faceted Browsing"
ASCOT (University of Washington)

- **Research in a Browser**
  - “iGoogle” for Astronomy
    - Collections of simple atomic applications (gadgets)
    - Users choose the view they want
    - All gadgets can communicate with each other
  - **Customizable and sharable**
    - Users can build and share “mashups”
    - Widgets are simple to create
    - Widgets call virtual observatory resources
  - **Efficient**
    - Communication is within the browser (fast)
    - Built from javascript (standard)
Select Gadgets

Rearrange based on your preference

Collaborative Astronomy (Connolly, Gibson, Krughoff, Sayers, Smith 2010)
All gadgets communicate through the data gadget

Name resolver and zoom to field

Query the SDSS based on viewport and return the source overlaid on images

Collaborative Astronomy (Connolly, Gibson, Krughoff, Sayers, Smith 2010)
Create, store and share multiple views of gadgets

Interaction allows selections to be shown on the viewport

Collaborative Astronomy (Connolly, Gibson, Krughoff, Sayers, Smith 2010)
A look at the editors: Jeff Dozier

This week, we spent some time with Jeff Dozier, professor in the Donald Bren School of Environmental Science & Management at the University of California, Santa Barbara, and learn more about what led him to his involvement with The Fourth Paradigm (first the book, now the blog). Here Jeff gives us a peek into his background starting with his love for mountain climbing, leading to his career as an educator, researcher and even the winner of the second annual Jim Gray eScience Award. With that, I’ll hand it over to Jeff...

*My path to a career as an academic scientist was not straightforward. About the time I graduated from high school, I stopped playing tennis and started rock climbing, and that was really my "major" in college. I had to leave school after a year-and-a-half because my grades were so bad. I dropped out, worked for a year, and made enough money to leave the country for 15 months, during which I studied German, climbed in England, Wales, Scotland, Poland, Austria, Italy, and France, and hitchhiked from Europe to India, from where I flew home. At the time (1968) I could not travel overland into Burma, so I flew there.

By then, I was 21 years old, realized that studying the Earth was what I wanted to do, and was able to judiciously mix school and climbing and do as well as I have summarizing climbing, almost having conditions in the heads.

http://blogs.nature.com/fourthparadigm/

http://sites.nationalacademies.org/PGA/brdi/index.htm
New, Seamless, Science?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. There are approximately 108,533,156,766 bases in 108,431,892 sequence records in the traditional GenBank divisions and 148,165,177,763 bases in 48,443,067 sequence records in the WGS division as of August 2006. The complete release notes for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

An example of a GenBank record may be viewed for a Saccharomyces cerevisiae gene.

The Centers for Disease Control and Prevention and other health officials are actively tracking the recent emergence of human cases of swine influenza A (H1N1) virus infection. Influenza A virus sequences from patients affected by this strain are being submitted to GenBank and can be accessed through the NCBI Flu Resource.
Seamless Astronomy
Alyssa A. Goodman
Harvard-Smithsonian Center for Astrophysics

with
Alberto Accomazzi, Douglas Burke, Gus Muench, Michael Kurtz & Alberto Pepe (Harvard-Smithsonian CfA); Eli Bressert (U. Exeter); Tim Clark (Massachusetts General Hospital/Harvard Medical School); Chris Borgman (UCLA); Jonathan Fay & Curtis Wong (Microsoft Research)