Albatross: rolling on a sea of data

Annette Bailey
Tracy Gilmore
Leslie O'Brien
Anthony Wright de Hernandez

Background

Evolution of usage stats management:

BUD, SUSHI, Scholarly Stats, 360 COUNTER, Foster

Considerations:

- Normalizing data
- Subscription costs vs personnel costs
- Flexibility of reporting

Big Ugly Database

Sponsor	Family / Platform	*warning -- don't sort below the data* Resource	YTD Total	Use Unit	Equiv vendor term	Data Entry By	Brief temp problem notes	Jul	Aug	Sep	Oct
VIVA	ABC-CLIO	America: History & Life	3,807	Searches	Searches	Metz		113	312	409	433
VIVA	ABC-CLIO	Historical Abs	1,462	Searches	Searches	Metz		9	35	240	201
VIVA	ACM	ACM	27,057	Documents	Total Articles	Metz		919	1,899	2,595	2,943
VIVA	ACS	American Chem Soc Journals	57,978	Documents	Article Requests	Metz		3,723	5,278	5,165	6,464
VT	ACS	SciFinder (Chem Abstracts)	73,433	Searches	Total Activity	Metz		4,682	6,821	8,179	8,129
VT	AIP	American Institute of Physics, ASCE, ASME, SPIE &many others (scitation)	27,069	Documents	Full Text Article Requests	Metz		3,283	2,367	2,117	2,448
VT	Alexander St	Oral History Online	1,760	Documents	Viewed Items	Metz		14	9	458	0
VT	Alexander St	Women & Social Mvmts	3,168	Documents	Viewed Items	Metz		485	31	401	8
VIVA	AMS	Mathscinet	33,466	Searches	Total Searches	Metz		2,813	2,729	2,596	2,772
VIVA	Annual Reviews	Annual Reviews	6,331	Documents	Full Text (HTML+PD	Metz		356	588	604	666
VT	Beilstein	Beilstein & Gmelin/Crossfire	3,148	Searches	Searches	Metz		44	389	307	365
VT	Biblioline (NISC)	Child Abuse, Child Welfare and Adoption	205	Searches	Searches	Metz	*be careful*	2	8	25	36
VT	Biblioline (NISC)	Family & Society Studies	2,234	Searches	Searches	Metz	display is	59	50	326	251
VT	Biblioline (NISC)	Fish & Fisheries	894	Searches	Searches	Metz	not quite	36	23	32	63
VT	Biblioline (NISC)	RILM (Music)	437	Searches	Searches	Metz	alpha in	2	43	26	34
VT	Biblioline (NISC)	Wildife & Ecology Studies	4,594	Searches	Searches	Metz	vendor	170	140	343	563
VT	Biblioline (NISC)	Women's Studies International	426	Searches	Searches	Metz	stats	9	24	62	37

Background

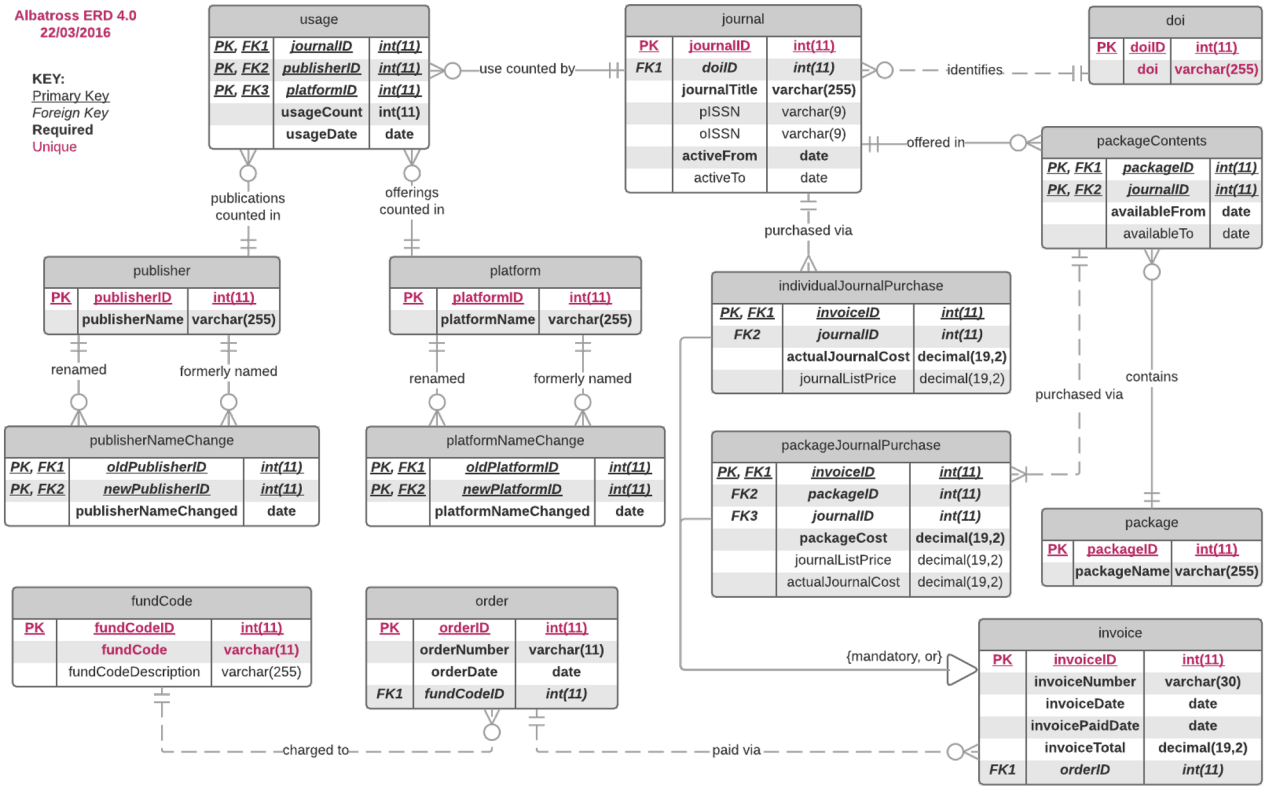
Evolution of usage stats management:

BUD, SUSHI, Scholarly Stats, 360 COUNTER, Foster

Considerations:

- Normalizing data
- Subscription costs vs personnel costs
- Flexibility of reporting

Navigating the Data: Entity Relationship Diagram



Navigating the Data: Entity Relationship Diagram

Positives

- Great automation
- Helpful for beginners
- Multiple platforms

Negatives

- User helps aren't that great
- Desktop software only
- License fees ≈ $100 / 3 mos.

Positives

- Free for Education
- Google Drive integration
- Web based

Negatives

- No automation
- Requires experience
- Requires Internet connection

Building the Database: Gathering Cost Data

Two types of cost data:

- List Price
  - Obtained from vendor lists
- Actual Cost
  - Obtained from the Sierra Database Navigator tool (SierraDNA).

Data Gathered

- 2,389 order rows
- 13,705 invoice rows

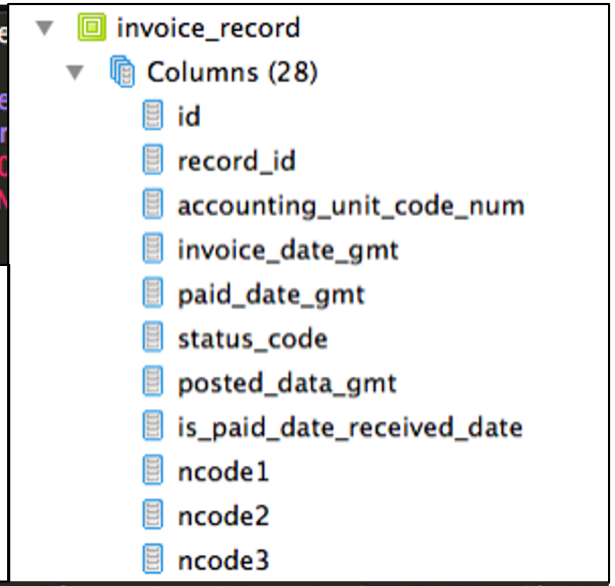
Allows comparison of

- paid price
- discounted (package) price
- list price

```
SELECT iv.invoice_number_text AS invoice
FROM sierra_view.invoice_view iv
    JOIN sierra_view.invoice_record_line
    JOIN sierra_view.order_view ov ON ir
    JOIN sierra_view.order_record orec
    JOIN sierra_view.form_property fp ON
    JOIN sierra_view.form_property_name
```

Costs pulled from invoice and order records required joining 6 tables to connect the DOI to the cost for each journal.

Not a lot - but it was challenging to identify the table connections in SierraDNA.



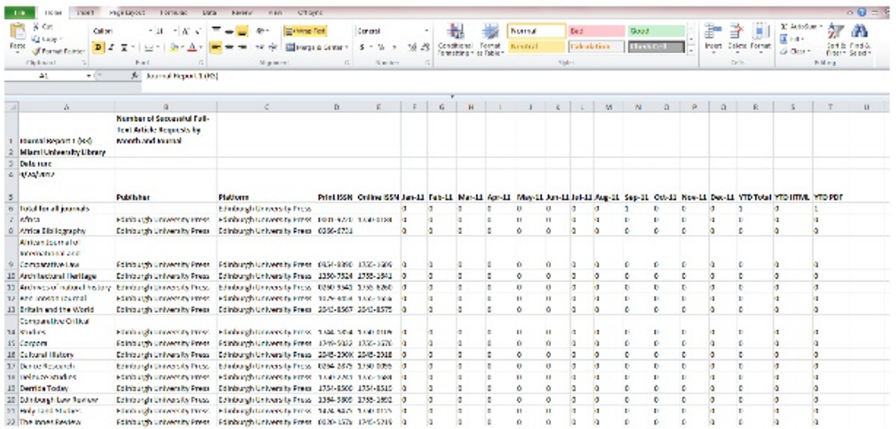

```
SELECT ov.record_num AS orderNumber, TO_CHAR(ov.order_date_gmt, 'MM/DD/YYYY') AS
FROM sierra_view.order_view ov
    JOIN sierra_view.order_record_cmf orc ON ov.record_id = orc.order_record_id
    JOIN sierra_view.fund_master fm ON CAST(orc.fund_code AS INT) = fm.code_num
    JOIN sierra_view.order_record orec ON ov.record_id = orec.id
    JOIN sierra_view.form_property fp ON orec.form_code = fp.code
    JOIN sierra_view.form_property_name fpn ON fp.id = fpn.form_property_id
```

Building the Database: Cleaning up JR1 COUNTER reports

Challenges working with 5 years of usage reports:

- Reports varied in structure and content
- Most reports were COUNTER release 3 or 4
- Many reports were non-COUNTER
- All reports required some manipulation
- 70% of reports both COUNTER and non-COUNTER required DOI's

COUNTER JR1 Report in Excel



Building the Database: Digital Object IDs

The majority of journal titles did not have DOIs

- Used CrossRef to find and add DOIs where available
- Created our own schema based on:
  - Journal Title initials
  - ISSN
  - 10.9999/ISSN or 10.9999/journal initials
- Identified over 100 journal packages and collections
  - Adding DOIs to titles
  - Identifying the list price

Building the Database: Not so standard data

Data from vendors - not following the standard closely - i.e. lack of Journal DOIs

COUNTER JR1 reports - no longer in standard format

We wrote scripts so that we can work across the cumulative data set. These scripts generate reports.

Building the Database: Publisher Data

```
 9 x '1105 Media, Inc.' (first occurrence ../data/2010/2010consolidated.xlsx:23309)
 1 x '1870-1930""' (first occurrence ../data/2012/2012consolidated.xlsx:2356)
 3 x '24/30 1/3 - Revue 24 images' (first occurrence ../data/2013/2013consolidated.xlsx:14280)
 1 x '3media Group Limited' (first occurrence ../data/2015/2015consolidated.xlsx:25172)
 1 x '4D Cross Connect.Com' (f
 1 x '580 Split' (first occurr
 3 x '828 E-Commerce International, Inc.' (first occurrence ../data/2010/2010consolidated.xlsx:25252)
 1 x '????' (first occurrence ../data/2014/2014consolidated.xlsx:10688)
 1 x '???? ??????' (first occurrence ../data/2015/2015consolidated.xlsx:30188)
 2 x '??????' (first occurrence ../data/2013/2013consolidated.xlsx:29990)
 1 x '??????-???????????????? ????? ????????????? ??????? ???????? ??? ???????' (first occurrence ../data/2015/2015consolidated.xlsx:28103)
 1 x '???????' (first occurrence ../data/2015/2015consolidated.xlsx:30002)
 1 x '???????? (??????????? ????????????????? ??????????? ???????? ? ????????)' (first occurrence ../data/2014/2014consolidated.xlsx:10415)
 3 x '????????' (first occurrence ../data/2013/2013consolidated.xlsx:3540)
 1 x '???????? ? ???????????? ???????????????? ????? ? ?????????' (first occurrence ../data/2015/2015consolidated.xlsx:25175)
 1 x '????????(???)???' (first occurrence ../data/2013/2013consolidated.xlsx:3545)
 3 x '?????????' (first occurrence ../data/2013/2013consolidated.xlsx:10958)
 1 x '??????????' (first occurrence ../data/2015/2015consolidated.xlsx:32725)
 1 x '?????????? ???????? ??????????' (first occurrence ../data/2013/2013consolidated.xlsx:9714)
 2 x '??????????? ????????????? ?????????? ????????????? ?????????? ??????? ???????????????? ????????' (first occurrence ../data/2013/2013consolidated.xlsx:10671)
 1 x '?????????????' (first occurrence ../data/2015/2015consolidated.xlsx:25329)
 2 x '??????????????' (first occurrence ../data/2013/2013consolidated.xlsx:10961)
 2 x '???????????????' (first occurrence ../data/2013/2013consolidated.xlsx:7615)
 3 x '?cole des Hautes ?tudes Commerciales de Montr?al' (first occurrence ../data/2010/2010consolidated.xlsx:25140)
12 x '?ditions Klincksieck' (first occurrence ../data/2010/2010consolidated.xlsx:21405)
 3 x '?scar Aguer' (first occurrence ../data/2010/2010consolidated.xlsx:26525)
 1 x 'A COLUMN BY ANGEL DE CAMPO (MICROS)""' (first occurrence ../data/2012/2012consolidated.xlsx:2207)
```

1 x '1870-1930""' (first occurrence ../data/2012/2012consolidated.xlsx:2356)
1 x '??????? (??????????? ????????????? ??????????? ???????? ? ????????)' (first occurrence ../data/2014/2014consolidated.xlsx:10415)

Building the Database: Journal Title Data

Rejecting title match between 'Work, Employment and Society' and 'Work, Employment & Society (Subscriber ID: 1052562)' despite similarity 0.450980
Mismatch (Work, Employment and Society,0950-0170,1469-8722) vs (Work, Employment & Society (Subscriber ID: 1052562),0950-0170,1469-8722)

Mismatch (Toxicologic pathology (Taylor & Francis),0192-6233,None) vs (Toxicologic Pathology (Subscriber ID: 1052562),0192-6233,1533-1601)
Rejecting title match between 'Work, Employment and Society' and 'Work, Employment & Society (Subscriber ID: 1052562)' despite similarity 0.450980
Mismatch (Work, Employment and Society,0950-0170,1469-8722) vs (Work, Employment & Society (Subscriber ID: 1052562),0950-0170,1469-8722)
Rejecting title match between 'Structures & Buildings' and 'Proceedings of the ICE - Structures and Buildings' despite similarity 0.428571

Rejecting title match between 'Computing in Science & Engineering (1999 -)' and 'Design & Test of Computers, IEEE (1985 -)' despite similarity 0.295455
Mismatch (Computing in Science & Engineering (1999 -),0740-7475,None) vs (Design & Test of Computers, IEEE (1985 -),0740-7475,None)

Rejecting title match between 'Acoustics, Speech, and Signal Processing Newsletter, IEEE (1974 - 1983)' and 'Audio, Transactions of the IRE Professional Group on (1953 - 1954)' despite similarity 0.338028
Mismatch (Acoustics, Speech, and Signal Processing Newsletter, IEEE (1974 - 1983),0000-0000,None) vs (Audio, Transactions of the IRE Professional Group on (1953 - 1954),0000-0000,None)
Rejecting title match between 'Component Parts, IRE Transactions on (1955 - 1962)' and 'Component Parts, Transactions of the IRE Professional Group on (1954 - 1955)' despite similarity 0.513158
Mismatch (Component Parts, IRE Transactions on (1955 - 1962),0096-2422,None) vs (Component Parts, Transactions of the IRE Professional Group on (1954 - 1955),0096-2422,None)
Rejecting title match between 'IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)' and 'Computational Biology and Bioinformatics, IEEE/ACM Transactions on (2004 -)' despite similarity 0.250000
Mismatch (IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB),1545-5963,None) vs (Computational Biology and Bioinformatics, IEEE/ACM Transactions on (2004 -),1545-5963,None)
Rejecting title match between 'Computing & Control Engineering Journal (1990 - 2007)' and 'Control & Automation (2007 -)' despite similarity 0.320755
Mismatch (Computing & Control Engineering Journal (1990 - 2007),0956-3385,None) vs (Control & Automation (2007 -),0956-3385,None)
Rejecting title match between 'Computing in Science & Engineering (1999 -)' and 'Design & Test of Computers, IEEE (1985 -)' despite similarity 0.295455
Mismatch (Computing in Science & Engineering (1999 -),0740-7475,None) vs (Design & Test of Computers, IEEE (1985 -),0740-7475,None)
Warning: ../data/2010/2010consolidated.xlsx:13374 has non-valid pISSN 1392-4553
Warning: ../data/2010/2010consolidated.xlsx:13375 has non-valid pISSN 1789-8348
Warning: ../data/2010/2010consolidated.xlsx:13376 has non-valid pISSN 0096-2053
Rejecting title match between 'IEE Proceedings -- Intelligent Transport System' and 'Intelligent Transport Systems, IEE Proceedings (2006 - 2006)' despite similarity 0.183333
Mismatch (IEE Proceedings -- Intelligent Transport System,1748-0248,None) vs (Intelligent Transport Systems, IEE Proceedings (2006 - 2006),1748-0248,None)
Rejecting title match between 'IEEE/ACM Transactions on Networking (TON)' and 'Networking, IEEE/ACM Transactions on (1993 -)' despite similarity 0.434783
Mismatch (IEEE/ACM Transactions on Networking (TON),1063-6692,None) vs (Networking, IEEE/ACM Transactions on (1993 -),1063-6692,None)

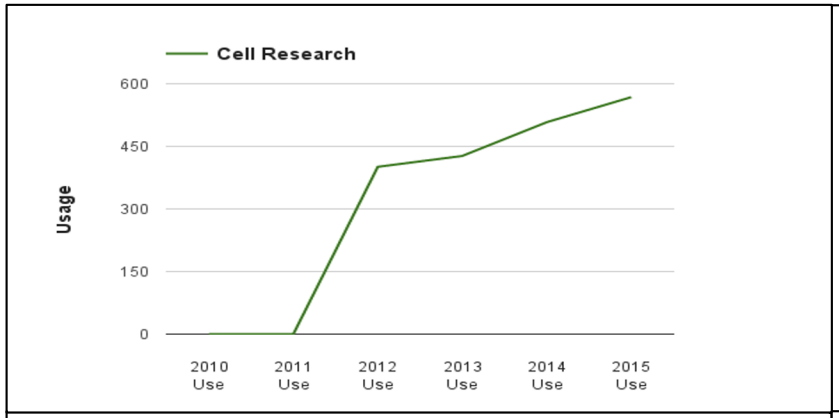
Using the Database: Workflows

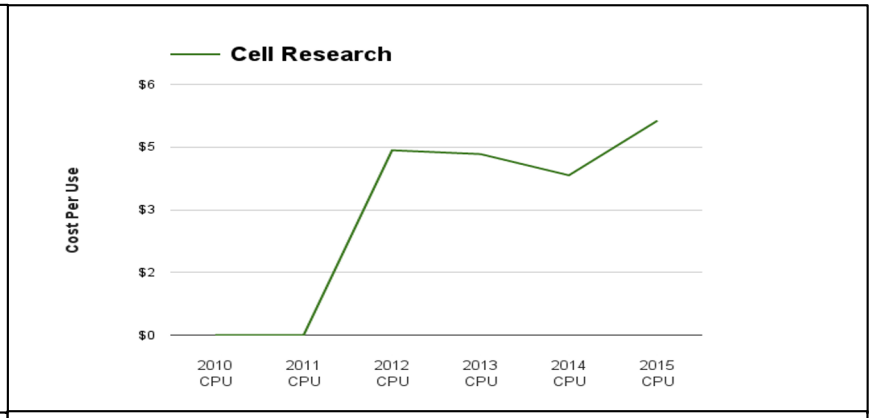
New skills and new tools were required to work with and visualize our data.

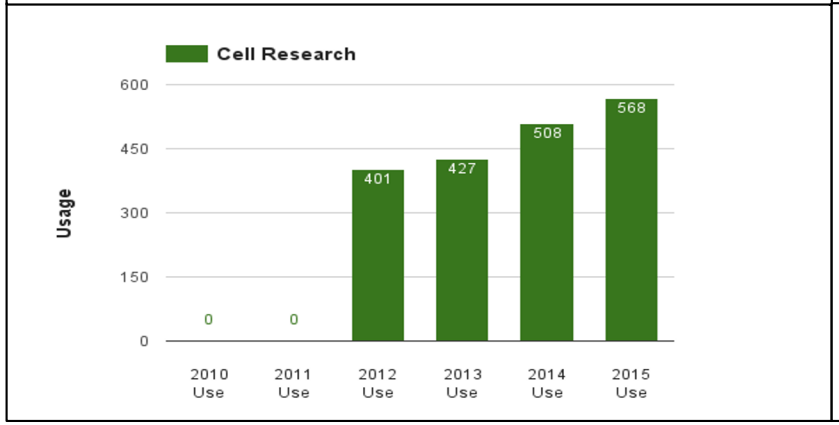
- SQL for querying the database
- VLookup for updating DOIs
- Pivot tables and Pivot charts
- Excel Dashboards
- Google Dashboards
- Tableau

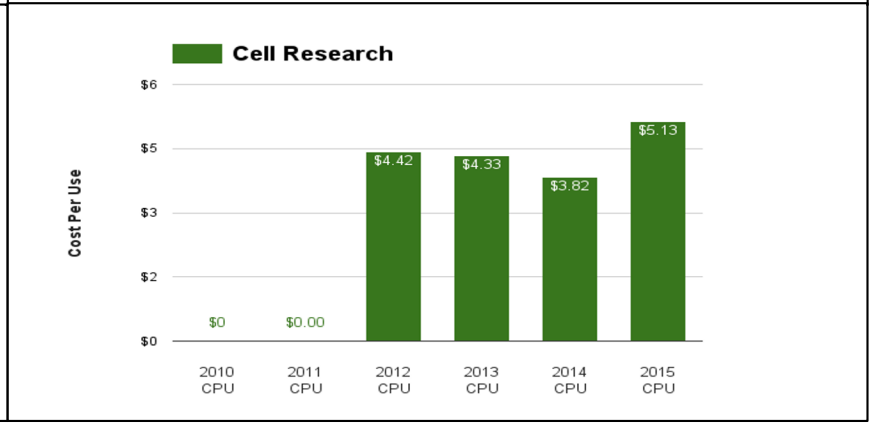
The purpose is to provide a more public and robust picture of our usage.

Using the Database: Reports









Why design our own database?

- Budget
- Control over the data
- Allows for robust reporting
  - Review of journal packages
  - Accommodates changes to COUNTER
- Flexibility and long-term sustainability

Evaluating the Database: Future Plans

- Expanding the data points
- Demonstrating the value to the university
- Evaluating package and big deals
- Automated download of COUNTER reports for maximum flexibility and less manual work
- Investigating data warehousing solutions and data visualization tools (in process):
  - Amazon Web Services
  - Google Drive
  - Hiring data visualization specialists

Questions?

Annette Bailey: afbailey@vt.edu
Tracy Gilmore: tgilmore@vt.edu
Leslie O'Brien: lobrien@vt.edu
Anthony Wright de Hernandez: antwri@vt.edu