Atlas: Data Sharing in HIV Research

Sarah Ramsay
CHAVI/VISC/CDS Program Manager, SCHARP
sramsay@scharp.org
Agenda

• Background
• Atlas data sources
• Atlas usage and value
• Challenges
• Lessons and recommendations
Background: SCHARP

• Provides statistical collaboration to infectious disease researchers around the world
  – Includes statistical methodology and mathematical modeling research
• Collects, manages, and analyzes data from clinical trials and epidemiological studies of infectious disease
• Part of the Vaccine and Infectious Disease Division (VIDD) of the Fred Hutchinson Cancer Research Center
• Funded as the Data Management and Statistical Center for 3 large HIV research networks (MTN, HVTN, HPTN)
  – CHAVI, CHAVI ID, CAVD and more
Background: Atlas

- Goal of increasing transparency and improving operational efficiency in distributed collaborations
- Development started in July, 2005, launched in August, 2006
- Primary contributors:

<table>
<thead>
<tr>
<th>Years</th>
<th>Network/Group</th>
<th>Funder</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005-2012</td>
<td>CHAVI</td>
<td>NIH (via Duke)</td>
<td>CRF sharing, specimen tracking</td>
</tr>
<tr>
<td>2006-2012</td>
<td>CAVD/VISC</td>
<td>BMGF</td>
<td>Assay tools: NAb, GPAT</td>
</tr>
<tr>
<td>2007-2008</td>
<td>SCHARP (Shared)</td>
<td>Multiple</td>
<td>Dev tools, APIs</td>
</tr>
<tr>
<td>2009-2012</td>
<td>HVTN</td>
<td>NIH</td>
<td>Admin features, study/specimen scalability</td>
</tr>
<tr>
<td>2009-2012</td>
<td>MTN</td>
<td>NIH</td>
<td>Full-text search</td>
</tr>
<tr>
<td>2011-2012</td>
<td>HPTN</td>
<td>NIH</td>
<td>Protocol-specific tools</td>
</tr>
</tbody>
</table>
Background: Atlas

• Originally began as part of the CHAVI and CAVD grants
  – Mandate for an online data sharing platform
• Large growth in the last 6 years of use
  – 2200 Active User Accounts
  – 3800 Folders
  – 46 Projects
Background: Atlas Usage

34,758 people visited this site

- Visits: 266,291
- Unique Visitors: 34,758
- Pageviews: 3,119,649
- Pages / Visit: 11.72
- Avg. Visit Duration: 00:10:28
- Bounce Rate: 18.31%
- % New Visits: 13.05%
- 86.93% Returning Visitor (231,499 Visits)
- 13.07% New Visitor (34,792 Visits)
Agenda

• Background
• Atlas data sources
• Atlas usage and value
• Challenges
• Lessons and recommendations
Data sources: overview

- **Atlas**
  - Lab and Assay
    - Luminex
    - NAb
    - “GPAT”
    - Etc.
  - Specimens
    - Lab and repository LIMS
    - Specimen requests
  - Clinical (CRF)
    - Demographics
    - Physical exams
    - Etc.
  - SCHARP
    - PDF reports
    - SAS datasets
    - SCHARP databases
    - SOPs and protocol docs
Data sources: clinical

- **DataFax**
  - SCHARP-run and maintained
  - 42 protocols, 53,781 forms in August
  - 118,394 forms to data processed and imported for the 7 CHAVI protocols
- **3 Pipelines to Atlas**
  - Datafax to Atlas
    - SCHARP-authored
    - Protocol-specific
    - Nightly import into Atlas Study Folders
    - Currently outputs a mix of TSV- and XML-based study formats
  - SAS to Atlas
    - SCHARP-authored
    - Nightly import into Atlas Study Folders
  - SAS Share
    - Direct external data source exposure in Atlas data grids
Data sources: clinical

Remote Sites

Clinical Site

CRF

Fax

Clinical Site

CRF

Fax

SCHARP

dfMirror

Legacy study archives

Atlas

SAS Share

SAS

XML-based study archives

DataFax
Data sources: Specimen

• Most Atlas specimen data starts with FSTRF
  – Data from > 132 location-specific LDMS installations is exported to FSTRF
  – FSTRF compiles and sends to SCHARP

• SCHARP-side pre-processing pipeline
  – Quality control checks
  – Data normalization
  – Exports to per-protocol or per-network LabKey Server specimen archives

• Data reloaded nightly into Atlas study folders
1. Data is exported from each location’s LDMS system to FSTRF and uploaded to SCHARP nightly.

2. Data is normalized and combined into a specimen archive. SCHARP has an extensive internally developed system for evaluating and integrating these data into a LabKey specimen archive.

3. The specimen archive is loaded into Atlas via a nightly job which is started automatically by the SCHARP-side processing pipeline.
**Data sources: Specimen**

<table>
<thead>
<tr>
<th>Summary (Vial Count)</th>
<th>Acute Cohort Enrollment</th>
<th>Int Vist</th>
<th>Acute Cohort Week 1</th>
<th>Int Vist</th>
<th>Acute Cohort Week 2</th>
<th>Int Vist</th>
<th>Acute Cohort Week 3</th>
<th>Int Vist</th>
<th>Acute Cohort Week 4</th>
<th>Int Vist</th>
<th>Acute Cohort Week  8</th>
<th>Int Vist</th>
<th>Acute Cohort Week 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood (Whole)</td>
<td>[unspecified]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blood (Whole)</td>
<td>208</td>
<td>186</td>
<td>180</td>
<td>14</td>
<td>212</td>
<td>6</td>
<td>218</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extracted DNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extracted RNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PBMC Cells, Viable</td>
<td>2213</td>
<td>1330</td>
<td>1354</td>
<td>1227</td>
<td>1445</td>
<td>1459</td>
<td>15</td>
<td>1703</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plasma</td>
<td>8878</td>
<td>5511</td>
<td>5616</td>
<td>5742</td>
<td>6265</td>
<td>6657</td>
<td>34</td>
<td>7546</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serum</td>
<td>803</td>
<td>618</td>
<td>627</td>
<td>621</td>
<td>766</td>
<td>723</td>
<td>3</td>
<td>736</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole Blood Pellet, Specify Methodology</td>
<td>325</td>
<td>118</td>
<td>122</td>
<td>144</td>
<td>125</td>
<td>164</td>
<td>149</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Breast Milk</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast Milk - Spun</td>
<td>18</td>
<td>22</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast Milk - Whole</td>
<td>8</td>
<td>13</td>
<td>3</td>
<td>15</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cryopreserved primary cells from Non-Blood Spec Type, Viable</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cervical Vaginal Lavage</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cryopreserved primary cells from Non-Blood Spec Type, Viable</td>
<td>30</td>
<td>1</td>
<td>25</td>
<td>1</td>
<td>33</td>
<td>30</td>
<td>36</td>
<td>2</td>
<td>28</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluid Portion from a Non-Blood Specimen Type</td>
<td>104</td>
<td>6</td>
<td>85</td>
<td>12</td>
<td>120</td>
<td>109</td>
<td>128</td>
<td>5</td>
<td>108</td>
<td>123</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cervix</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cryopreserved primary cells from Non-Blood Spec Type, Viable</td>
<td>[unspecified]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluid Portion from a Non-Blood Specimen Type</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tear Flo Strips</td>
<td>25</td>
<td>1</td>
<td>19</td>
<td>2</td>
<td>27</td>
<td>22</td>
<td>25</td>
<td>1</td>
<td>19</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Saliva</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saliva</td>
<td>73</td>
<td>63</td>
<td>71</td>
<td>70</td>
<td>87</td>
<td>57</td>
<td>4</td>
<td>65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Semen</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cryopreserved primary cells from Non-Blood Spec Type, Viable</td>
<td>[unspecified]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluid Portion from a Non-Blood Specimen Type</td>
<td>54</td>
<td>80</td>
<td>90</td>
<td>2</td>
<td>84</td>
<td>85</td>
<td>2</td>
<td>92</td>
<td></td>
<td>99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Urine</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urine</td>
<td>1498</td>
<td>1200</td>
<td>1240</td>
<td>1165</td>
<td>1390</td>
<td>253</td>
<td>16</td>
<td>262</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[unspecified]</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data sources: Assay

- Assay data is the most diverse in format and origin
  - >7 labs, >6 assay types
  - Many more labs deliver data via file drop (30+)
  - New data formats appear monthly
- Direct lab import
  - NAb, Luminex, GPAT data uploaded directly to lab folders
  - “Copy to study” pushes QC’d data into shared folders
- Indirect/assisted import (via LDO)
  - Data file drops (FTP, Atlas file management tools)
  - Email
  - Other custom tools external to Atlas
Data sources: Assay

- Lab A
- Lab B
- Lab C

Data entry tools:
- Atlas GPAT tool
- Atlas NAb tool
- Atlas Luminex tool
- Atlas file upload
- File drop (FTP)
- Email

Sharing via Atlas

- Copy to study
- Atlas extract

LDO

SAS

- Manual upload
- SAS Share or study import
Data sources: Assay

- 3 Tiers of Assay Data
  - Standardized (and sometimes validated) has a well defined structure and analysis plan
  - Non-standardized usually has a good structure, but may change depending on developing analysis trends
  - R&D, highly unstable data structures with quick changes to layout and analysis
Data sources: NAb example

Run Summary: NAb1

- Assay Id: NAb1
- Created: 2009-09-03 15:06:30.354
- Created By: eknelson
- Virus Name: HIV-1
- Virus ID: P392
- Host Cell: T
- Study Name: Demo
- Experiment Performer: Elizabeth

- Experiment ID: Nab32
- Incubation Time: 30
- Plate Number: 1
- File ID: NAbresults2
- Lock Graph Y-Axis: false
- Curve Fit Method: Five Parameter
- Batch: 2009-09-03 batch 2

Range: 54871
Virus Control: 55931 ± 6%
Cell Control: 1060 ± 6%

Cutoff Dilutions
<table>
<thead>
<tr>
<th>Curve Based</th>
<th>Point Based</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>80%</td>
</tr>
<tr>
<td>526455390.2504.346</td>
<td>117</td>
</tr>
<tr>
<td>249325717.2404.493</td>
<td>504</td>
</tr>
<tr>
<td>249320619.2604.640</td>
<td>832</td>
</tr>
<tr>
<td>249328595.2604.530</td>
<td>979</td>
</tr>
<tr>
<td>526455350.4404.456</td>
<td>425</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Specimen ID</th>
<th>Participant ID</th>
<th>Visit ID</th>
<th>Date</th>
<th>Dilution Factor</th>
<th>Initial Dilution</th>
<th>Method</th>
<th>Fit Error</th>
<th>AUC</th>
<th>PositiveAUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>526455390.2504.346</td>
<td>249325717.2404.493</td>
<td>2804.0</td>
<td>2008-09-02</td>
<td>3.0</td>
<td>20.0</td>
<td>Dilution</td>
<td>4.7</td>
<td>0.224</td>
<td>0.238</td>
</tr>
<tr>
<td>249325717.2404.493</td>
<td>249320107</td>
<td>2804.0</td>
<td>2008-07-30</td>
<td>3.0</td>
<td>20.0</td>
<td>Dilution</td>
<td>4.1</td>
<td>0.372</td>
<td>0.405</td>
</tr>
<tr>
<td>249320619.2604.640</td>
<td>249320127</td>
<td>2804.0</td>
<td>2009-01-06</td>
<td>3.0</td>
<td>20.0</td>
<td>Dilution</td>
<td>3.4</td>
<td>0.477</td>
<td>0.477</td>
</tr>
<tr>
<td>249328595.2604.530</td>
<td>249320489</td>
<td>2804.0</td>
<td>2008-12-03</td>
<td>3.0</td>
<td>20.0</td>
<td>Dilution</td>
<td>4.5</td>
<td>0.547</td>
<td>0.547</td>
</tr>
<tr>
<td>526455350.4404.456</td>
<td>249320897</td>
<td>2804.0</td>
<td>2008-07-20</td>
<td>3.0</td>
<td>20.0</td>
<td>Dilution</td>
<td>2.9</td>
<td>0.378</td>
<td>0.390</td>
</tr>
</tbody>
</table>
Data sources: SCHARP-produced

• Atlas is also used to share data from internal systems
  – SAS datasets (via SAS Share)
  – SOPs, Protocol documents
  – Analysis results in various formats
  – Lists of antibodies, virus isolates, isotypes, etc.
  – Other postgres databases

• Permissions vary
  – Facilitating both internal and external workflows
Data sources: SCHARP-produced

- USMHRP RV144 Correlates Analysis
  - Study Management system
- Adjudication Tools
- Assay Monitoring
- Specimen DB
HIVIG, all viruses, performed by everyone, in all studies, all dates

All Matching Results [Export to Excel]

<table>
<thead>
<tr>
<th>Study</th>
<th>Virus Name</th>
<th>Experiment Date</th>
<th>Performed By</th>
<th>Curve ID50</th>
</tr>
</thead>
<tbody>
<tr>
<td>[run details]</td>
<td>Haynes_McCormik_Zambia</td>
<td>SVA-MLV</td>
<td>2009-02-05</td>
<td>Forst, Kathleen</td>
</tr>
<tr>
<td>[run details]</td>
<td>Haynes_McCormik_Zambia</td>
<td>6535.3</td>
<td>2009-02-05</td>
<td>Forst, Kathleen</td>
</tr>
<tr>
<td>[run details]</td>
<td>Haynes_McCormik_Zambia</td>
<td>QH0692.42</td>
<td>2009-02-05</td>
<td>Forst, Kathleen</td>
</tr>
<tr>
<td>[run details]</td>
<td>Haynes_McCormik_Zambia</td>
<td>SC422661.8</td>
<td>2009-02-05</td>
<td>Forst, Kathleen</td>
</tr>
</tbody>
</table>
**Most Recent Pipeline Job Status**

**Job Initiated At:** 2012/09/15 19:00:02

- **Parse LDMS-CHAVI:** 31 records were excluded from import.
- **Parse Labware-CHAVI:** 1260 records were excluded from import.
- **Parse LDMS-MTN:** 1785 records were excluded from import.
- **Parse LDMS-HPTN:** 1661 records were excluded from import.
- **Import Manifest Specimens:** 24 records were excluded from import.
- **Post Import Processing:** The number of vials imported for lab 121 is significantly larger than usual.
- **Export Data:** The number of vials exported for lab 121 is significantly larger than usual.

**File Size (kB):**

<table>
<thead>
<tr>
<th>File Size (kB)</th>
<th>LDMS-CHAVI</th>
<th>LDMS-HVTN</th>
<th>LDMS-MTN</th>
<th>LDMS-HPTN</th>
<th>Labware-CHAVI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vials Imported</td>
<td>336,966</td>
<td>337,859</td>
<td>132,271</td>
<td>555,937</td>
<td>11,384</td>
</tr>
<tr>
<td>Vials Exported</td>
<td>577,452</td>
<td>963,513</td>
<td>399,391</td>
<td>1,338,886</td>
<td>84,428</td>
</tr>
<tr>
<td>File Date/Time</td>
<td>2012/09/16</td>
<td>2012/09/16</td>
<td>2012/09/16</td>
<td>2012/09/16</td>
<td>2012/09/16</td>
</tr>
<tr>
<td>Upload Time</td>
<td>50m 30s</td>
<td>56m 22s</td>
<td>23m 5s</td>
<td>1h 25m 59s</td>
<td>3m 56s</td>
</tr>
</tbody>
</table>

**Job Initiated At:** 2012/09/16 19:00:02

**Job Completed At:** 2012/09/16 22:59:39

**SCHARP Pipeline Time:** 3h 59m 37s
Agenda

• Background
• Atlas data sources
• **Atlas usage and value**
• Challenges
• Lessons and recommendations
Value to external collaborators

- **Funders**
  - Real-time reports available to DAIDS (e.g., HVTN IQC/EQC)

- **Public**
  - Data from ~70 CAVD studies available for public access
  - Free download of documents for 16 HPTN and 15 MTN protocols
Value to external collaborators

• Network core
  – Transparency into data management and analysis

• Labs
  – Access to CRF (clinical) data
  – Rapid access to integrated reports (e.g., Borrow queries)
  – Operational efficiencies when Atlas is workflow-integrated (e.g., NAb)
  – Example: cross-comparison of CRF and specimen data allowed automatic identification of mislabeled CHAVI vials
Atlas value to SCHARP staff

• Primary Investigators
  – Atlas is a selling point in grants/proposals

• SCHARP Operations
  – Secured delivery of reports (DSMB)
  – structured data delivery
  – Embedded quality control during data upload
    • Ptid, visit, specimen checks

• Statisticians
  – Flexible access (R, SAS, Excel) to data
Agenda

- Background
- Atlas data sources
- Atlas usage and value
- Challenges
- Lessons and recommendations
Challenges: people, politics and culture

• New drive towards collaborative research
  – HIV field historically competitive
  – Concerns that data will be misinterpreted

• Those paying the costs may not see the benefits
  – Technicians and project managers work to import data
  – PIs and statisticians benefit from growing data asset

• Change is difficult
  – Ownership of current methodologies hinders progress
  – Resistance to change is independent of need for change
Challenges: technological

• Flexibility is a double-edged sword
  – Rapid tool development is possible and desirable
  – Long-term support of vast tool library is very expensive
• In-house expertise in the LabKey platform
  – Difficult to hire staff with LabKey experience
• Hard to gauge technical path forward given the variety of available resources
Agenda

- Background
- Atlas data sources
- Atlas usage and value
- Challenges
- Lessons and recommendations
Lessons learned: platform development

• Interaction with LabKey
  – early on we were too hands off: features can miss the mark
  – Over correction to intense oversight: expensive, slow iteration
  – Work to achieve balance in team based approach

• Don’t underestimate support costs
  – Allowed for organic growth of the system with downstream effects
  – “Quick” or “small” tools can be expensive
  – Documentation, standardization, planning are needed
  – Acceptance/regression testing

• Invest in self-empowerment
  – Developer tools/APIs and administrative tools have paid for themselves many times over
Lessons learned: platform adoption

• Lab buy-in requires covering their full workflow
  – Show immediate value after upload

• Adoption takes time
  – Much faster if system is integral to users’ daily work
Any questions?

Sarah Ramsay
sramsay@scharp.org
(206) 667-6992