Science Panel Discussion presentation: "You Want to Do What? Managing and Distributing Identifying Data without Running Afoul of Your Research Sponsor, Your IRB, or Your Office of Counsel"

Patrick J. Flynn
University of Notre Dame

Follow this and additional works at: http://escholarship.umassmed.edu/escience_symposium

Part of the Computer Sciences Commons, Engineering Commons, and the Library and Information Science Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License.
“You want to do WHAT?”

Managing and distributing identifying data without running afoul of your research sponsor, your IRB, or your Office of Counsel

Patrick J. Flynn
Dept. of Computer Science and Engineering
flynn@nd.edu
Outline

- Context: Sponsored projects to support biometric data sampling
- General approach to sampling
- Human Subjects issues
- Distributing the data
- Infrastructure: How to manage all of that data internally so you can find and fix errors
- Concluding remarks
Context

- ND has been funded since 2001 to conduct basic research in biometrics
- We collect biometric samples to support US Govt evaluations and foster increased R&D in biometrics
- Large project, many undergraduate and graduate students, postdocs, staff
- Funding mix mostly US Govt, some industry
General Approach to Sampling

- Sponsor and ND agree on subject count per week, # of weeks, and sensors in use
- “Subject session”: all data collected from a subject when they show up in the lab during a week of acquisitions
- Size of “subject session” has increased DRAMATICALLY with time...
- ... hence, overall data “burden” has increased (collecting 50-100 GB per acquisition week, 10 weeks/year)
>50 GB per week... wat?

- We collect images and videos of people's faces and irises and scripted “activities” using many consumer and specialized cameras.
An acquisition week

- Collect data for 18 hours spread over Tuesday-Thursday
- Three locations (studio and two other locations with uncontrolled illumination)
- 50 undergraduate operators, 13 GRA operators, etc.
- End-of-day backup of data from cameras and host computers to secure storage
After an acquisition week... enrollment.

- Ingest data into “BXgrid” portal/database/distributed filer system
- Dozens of “validation” assignments to operators
- After validation, some image and video data and associated metadata is delivered to USG sponsor for placement in a research archive
Data release and licensing

- Sponsor often “embargoes” data from wider release until programmatic use is complete.
- Then, we are authorized to design data sets for release to authorized research groups.
- Currently 18 such data sets, >1000 active licenses.
We're capturing large sets of pictures of people, giving them to the US Government and licensing them to other research groups.

Q. What could go wrong?

A. Lots could go wrong... so we are very careful about collecting and distributing data.
Human Subjects IRB

- Standing committee (Federal mandate) that approves all requests for research involving human subjects
- Required: a protocol that has certain required elements (hypothesis, subject pool, privacy provisions, risks and benefits, ...)
- Key element is the Consent Form
Consent form

- Legal record of agreement to participate
- Stylized format with required elements that summarize the protocol
- Signature required
- Must be kept for several years (i.e., forever)
IRB Approval

- Protocol & consent form must be approved by IRB before research (data collection) begins
- It is NOT a formality... protocols are scrutinized very carefully (I'm on ND’s IRB)
- Need to give yourself 2 months' lead time for ND approval
- Expedited reviews and exemptions possible (but IRB has to approve these also)
Even more IRB approval!

- Assertion: US agencies sponsoring research are getting much more picky about compliance over time

- Example: our data collections were sent out to the US Army for approval after ND approval was obtained.
  - ND approval: 2 days (expedited)
  - Army approval: 3 months

- NIST also is installing new IRB requirements
Who can get the data?

- Unrestricted USG by mandate
- Everyone else must license the data
- We have a standard license that we worked out with ND’s Office of Counsel in 2002 and have used with minor changes ever since.
Counsel involvement

We were proactive in our interactions with Counsel

Our perspective
- we need education on privacy
- understand ND-specific considerations
- Counsel’s perspective: facilitating scholarly inquiry, but be sensitive to risk

MANY contacts over the years (new collections, license negotiations, etc.)
License terms

- No redistribution
- No commercial use (carefully defined)
- Publication requirements (if reproducing >10 images, must ask permission), don't use ND's name to market your stuff, don't embarrass the subject
- Citation to an appropriate ND-authored publication about the data
- Copies of publications to ND
- Indemnification
Infrastructure

- We needed a system for organizing/managing data.
  - Ingest data and metadata
  - Fix the inevitable errors
  - Make it “easy” to do both
- Database & Web & storage technologies are key elements (the latter two are CS research areas in their own right)
Infrastructure

- Bxgrid: NSF-supported, database-backed, distributed-file-storage-using, Web-front-ended biometric data management system
- LOTS of data in it (but not not all of our data)

<table>
<thead>
<tr>
<th>Browse Source Datasets:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Face Images (bxgrid)</td>
<td>367368 records</td>
</tr>
<tr>
<td>2. Face Videos (bxgrid)</td>
<td>38944 records</td>
</tr>
<tr>
<td>3. Iris Images (bxgrid)</td>
<td>474295 records</td>
</tr>
<tr>
<td>4. Iris Videos (bxgrid)</td>
<td>11753 records</td>
</tr>
<tr>
<td>5. 3D Face Scans (bxgrid)</td>
<td>18410 records</td>
</tr>
<tr>
<td>6. Environments (bxgrid)</td>
<td>88 records</td>
</tr>
<tr>
<td>7. Illuminants (bxgrid)</td>
<td>28 records</td>
</tr>
<tr>
<td>8. Sensors (bxgrid)</td>
<td>57 records</td>
</tr>
<tr>
<td>9. Stages (bxgrid)</td>
<td>73 records</td>
</tr>
<tr>
<td>10. Subjects (bxgrid)</td>
<td>3218 records</td>
</tr>
<tr>
<td>11. Collections (bxgrid)</td>
<td>66 records</td>
</tr>
<tr>
<td>12. Metadata Log (bxgrid)</td>
<td>4805092 records</td>
</tr>
</tbody>
</table>
### BXgrid demo

#### BXGRID - BIOMETRICS RESEARCH GRID

Logged in as: flynn

Showing 1 to 10 of 942 results. Download all results as **TXT** or **CSV** or **XML** or **TAR**.

<table>
<thead>
<tr>
<th>thumbnail</th>
<th>recordingid</th>
<th>subjectid</th>
<th>collectionid</th>
<th>sensorid</th>
<th>eye</th>
<th>color</th>
<th>glasses</th>
<th>gender</th>
<th>YOB</th>
<th>race</th>
<th>date</th>
<th>fileid</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="nd3R53192" alt="image" /></td>
<td>nd3R53192</td>
<td>nd1S04261</td>
<td>nd1C00005</td>
<td>nd1N00006</td>
<td>Left</td>
<td>Hazel</td>
<td>No</td>
<td>Male</td>
<td>1964</td>
<td>White</td>
<td>2004-08-31 13:00:00</td>
<td>1602</td>
</tr>
<tr>
<td><img src="nd3R53191" alt="image" /></td>
<td>nd3R53191</td>
<td>nd1S04261</td>
<td>nd1C00005</td>
<td>nd1N00006</td>
<td>Left</td>
<td>Hazel</td>
<td>No</td>
<td>Male</td>
<td>1964</td>
<td>White</td>
<td>2004-08-31 13:00:00</td>
<td>1602</td>
</tr>
<tr>
<td><img src="nd3R53190" alt="image" /></td>
<td>nd3R53190</td>
<td>nd1S04261</td>
<td>nd1C00005</td>
<td>nd1N00006</td>
<td>Left</td>
<td>Hazel</td>
<td>No</td>
<td>Male</td>
<td>1964</td>
<td>White</td>
<td>2004-08-31 13:00:00</td>
<td>1602</td>
</tr>
</tbody>
</table>
Conclusions

- Collecting lots of research data presents challenges
- If it's human subject data, more challenges
- If it relates to identity, even more challenges!
- Critical elements for success:
  - Know thy context (collect data correctly)
  - Know thy customers (sponsors, licensees)
  - Know thy constraints (IRB, Counsel)
  - Create and maintain infrastructure to manage the data
  - Develop a “culture of stewardship”
Thank you!

Contact: flynn@nd.edu