Purposeful gaming and BHL: engaging the public in improving and enhancing access to digital texts

William Ulate R.
Global BHL Coordinator
BHL-US/UK Technical Director

Florida, December 16 – 20, 2013
Project objectives and benefits

• Test new means of crowdsourcing to support the enhancement of content in BHL

• Demonstrate if digital games are an effective tool for analyzing and improving digital outputs from OCR and transcription

• Benefits of gaming include:
  – improved access to content by providing richer and more accurate data;
  – an extension of limited staff resources; and
  – exposure of library content to communities who may not know about the collections otherwise.
Purposeful Gaming and BHL

Digitalkoot

Missouri Botanical Garden
Harvard University
Cornell University
New York Botanical Garden
Purposeful Gaming and BHL
Purposeful Gaming and BHL
Purposeful Gaming and BHL

SCORE: 12345
SAVED: 4/10
Purposeful Gaming and BHL
Purposeful Gaming

Workflow Diagram for Processes and Decision Points

1. Digitize
   - Generate OCR 1 (Internet Archive)
   - Automated Quality Assessment

2. Clean-up BHL corpus
   - Fine-tune Text Correction Process
   - Evaluate accuracy scores for a sample of assembled corrected text
   - Assemble corrected text output
   - Add to Ground Truth files
   - Manual Correction

3. Run Game to Gather Corrections
   - Push Differences to the Game
   - Acceptable accuracy score?

4. Generate OCR 2 (MOBOT)
   - > 20% non-alpha characters?
   - > 1 OCR output?

5. Transcribe (2 outputs)
   - YES
   - Transcribe
   - Automated Text Comparison
   - Manual Review
   - NO

6. Discard?
   - YES
   - Mark as Discarded
   - NO
German text interpreted by the OCR process as:

“unb auf ben Â©elnrgen be6 fublic{}en”
**OCR Improvements**

<table>
<thead>
<tr>
<th></th>
<th>IA OCR</th>
<th>OCR 2</th>
<th>Transcription 1</th>
<th>Transcription 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>unb</td>
<td>und</td>
<td>und</td>
<td>und</td>
</tr>
<tr>
<td>2</td>
<td>den</td>
<td>ben</td>
<td>den</td>
<td>den</td>
</tr>
<tr>
<td>3</td>
<td>Å©elnrgen</td>
<td>Å©ebirgen</td>
<td>Bebirgen</td>
<td>Gebirgen</td>
</tr>
<tr>
<td>4</td>
<td>be6</td>
<td>des</td>
<td>de5</td>
<td>des</td>
</tr>
<tr>
<td>5</td>
<td>fublic{}en</td>
<td>fublichen</td>
<td>Füdlichen</td>
<td>Südlichen</td>
</tr>
<tr>
<td>6</td>
<td>Å£)eittfc{}(anb6</td>
<td>Deutschlanbs</td>
<td>Deutschlands</td>
<td>Deutschlands</td>
</tr>
</tbody>
</table>

Different resulting texts from parsing the phrase: “und auf den Gebirgen des südlichen Deutschlands”

(“and on the mountains of southern Germany”)
Purposeful Gaming

Workflow Diagram for Processes and Decision Points

1. **Digitize**
   - Generate OCR 1 (Internet Archive)
   - Automated Quality Assessment

2. **Clean-up BHL corpus**
   - Fine-tune Text Correction Process
   - Evaluate accuracy scores for a sample of assembled corrected text
   - Assemble corrected text output
   - Add to Ground Truth files
   - Manual Correction
   - Mark as Discarded

3. **Generate OCR 2 (MOBOT)**
   - > 20% non-alpha characters?
   - > 1 OCR output?
   - Transcribe (2 outputs)
   - Run Game to Gather Corrections
   - Push Differences to the Game
   - Acceptable accuracy score?
   - Manual Review

4. **Discard?**

Flowchart shows decision points and processes involved in the workflow of purposeful gaming.
2013 iDigBio Transcription Hackathon
December 16 - 20, 2013

Thank you

Trish.Rose-Sandler@mobot.org
BHL Data Analyst
Purposeful Gaming Project P.I.
Missouri Botanical Garden
Trish.rose-sandler@mobot.org

William.Ulate@mobot.org
Global BHL Project Manager
BHL Technical Director
Missouri Botanical Garden
william.ulate@mobot.org