IMLS Interim Narrative Report (Dec 1, 2013-Nov 30, 2014)

Project Title: Purposeful Gaming and BHL: engaging the public in improving and enhancing access to digital texts

This narrative report outlines our accomplishments and challenges during year 1 of our 2 year grant. We are proud to report we are on track and within budget of our committed objectives.

Activities Completed

Objective 1: Digitizing Horticultural catalogs

Before digitization could begin, project partners needed to determine their selection criteria. The task of identifying appropriate catalogs to scan was complicated by efforts to minimize duplication of the project partners digitizing initiatives, but important in making the best use of the limited funding available for scanning. A task force for comparing holdings records was formed composed of a member from each project partner responsible for digitization - Missouri Botanical Garden (MOBOT); Cornell; and New York Botanical Garden (NYBG). Cornell took the lead in collecting and analyzing spreadsheets of all the collaborator's holdings. During this time, partners became aware that the USDA National Agricultural Library (NAL) had digitized a significant number of seed and nursery catalogs and that they were available within the Internet Archive (IA). Although not originally named as a partner on the grant, it seemed prudent to ask NAL to join the taskforce and include their records as part of the comparison process. The task force identified where holdings overlapped and which institutions had unique items, thereby highlighting which items to select for digitization.

These catalogs were digitized and assessed for quality and are now available within both IA and BHL. OCR for each page was generated as part of the upload process to IA. They include 122 items (18k pages) from MOBOT and 154 items (21k pages) from NYBG, with the remaining 11,812 items (556k pages) coming from NAL. In order to aggregate these materials, BHL created a Seed & Nursery Catalog collection http://biodiversitylibrary.org/browse/collection/seedcatalogs. Cornell catalogs are not yet available within BHL because their digitization is outsourced. They have shipped approximately 350 items (15k pages) to their digitization vendor with another 15k in the queue.

Objective 2: Transcribing notebooks and catalogs

Project partners at Harvard took the lead on researching and implementing transcription software. They formed a task force to facilitate selection of a tool for the project with an eye toward development of a long-term transcription solution for BHL. Multiple open source transcription tools were investigated including: 1) DigiVol (Atlas of Living Australia Biodiversity Volunteer Portal) 2) FromThePage 3) Scripto 4) Smithsonian Transcription Tool 5) T-PEN 6) Transcribe Bentham and 7) Transcribr. The decision was narrowed down to DigiVol and FromThePage as these tools best met the requirements of the project. There were no tools that supported multiple transcriptions of a single page, so it was decided to implement both tools to provide two transcriptions per page as required for the game. Additionally, it was advantageous that DigiVol came with an existing community of transcribers.

In May of 2014, 10 volumes of William Brewster's ornithological field notebooks (five diaries and five journals amounting to approximately 2000 pages) were uploaded to DigiVol and FromThePage. DigiVol volunteers immediately began transcribing Brewster's content. A project assistant at Harvard was hired to promote the transcription work and the overall project through traditional and social media outlets (blogging, Twitter, etc) – see media coverage in Objective 8. Two library assistants in the Ernst Mayr Library spend approximately 12 hours per week transcribing. As of November 24, 2014, 65% of the pages in DigiVol have been transcribed, and 25% of the pages in FromThePage have been transcribed. Once the Brewster materials are complete, horticultural catalogs whose OCR was problematic will be pushed to these tools for transcription.

Objective 3: Building the technical framework

For this project we proposed correction of two primary types of texts – 1) published texts including books, journals and horticultural catalogs and 2) hand written field notes. Digital outputs from each text page will include either an OCR output or a transcription depending on whether it has to be automatically or manually interpreted. Each output requires different processes and decision points in order to determine if it is a good candidate for the game (see Workflow Diagram in Appendix 1). A technical framework was needed to manage both the various outputs as well as manage the status for a page at each stage of the workflow. The team at MOBOT decided to adopt a free, open source tool called MediaWiki (www.mediawiki.org) for managing the final corrected texts and an open-source document database called mongoDB (http://www.mongodb.org/) for managing the status and intermediate outputs for each page in the workflow.

The team also developed an algorithm to automatically assess the quality of the OCR outputs to be able to determine which pages should be sent through the transcription step. Initially, staff hypothesized that a basic 80/20 rule would apply (i.e. if the OCR contained more than 20% of non-alpha characters it would not a suitable candidate for OCR and there would need to be manually transcribed). Upon testing of this hypothesis though it was determined that that measuring alpha-numeric character ratios did not provide a useful measure of OCR quality. However, an alternate algorithm, called Darwin-Core (https://github.com/idigbio-citsci-hackathon/darwin-score) was identified and adapted. This algorithm compares each document against dictionaries and a score is assigned based on the number of recognized words in the document. Those scores provide a way to identify candidates for transcription. MOBOT staff retained the original dictionaries that came with the script and added additional dictionaries that were more relevant to published texts.

Objective 4: Comparing digital outputs for accuracy

Once two digital outputs are generated for each page they have to be compared and given an accuracy score. If the score is acceptable those words with differences will be pushed to the game for verification by the players. If the score is not acceptable, the page will either be discarded or run through a manual correction process. MOBOT staff developed a preliminary text comparison algorithm. While somewhat rough, it is being refined by the removal of noise characters such as leading or trailing non-alphanumeric characters (e.g. "platypus" and "platypus." type differences will not be sent to the game).

Words that are pushed to the game will need both the digital outputs for that word as well as the coordinate boundaries so that a picture of the word can be extracted and presented to game players for verification. This is fairly easy for OCR outputs but not as easy for transcription outputs (see Findings/Accomplishments section)

Inputs for the gaming platform will be delivered in a lightweight data interchange format known as JSON (www.json.org), an open standard format. MOBOT staff sent multiple sample input files to the gaming platform to ensure compatibility.

Objective 5: Developing and Deploying game

Before game development could begin, a game designer had to be identified and hired. MOBOT began this process by compiling a list of appropriate gaming companies which came from a mix of recommendations, local design contacts, and searching for computing and gaming labs located within universities. The Request for Proposal (RFP) was written in March of 2014. Not wanting to become inundated with irrelevant responses that would not meet our needs, we made a conscious decision not to send out the RFP widely but rather targeted specific organizations on our list. We had a small budget and felt that the interest in our proposal would be somewhat limited to non-profit/research based game labs. We sent out the RFP to 12 organizations in April and received responses by May. Although many companies expressed interest in reviewing our RFP, in the end we only received 3 proposals which was less than we had hoped. Fortunately, this did not prove to be a problem because at least two of the proposals were very viable. For reviewing the proposals and choosing the gaming company, we put together a taskforce composed of members from the project team and a BHL representative who had game design experience. The taskforce selected the winner, Tiltfactor from Dartmouth (http://www.tiltfactor.org/) in July. Tiltfactor had several strengths that made their bid stand out, including extensive experience designing games for the education sector and extensive research into the psychological impact of those designs on players.

Discovery and design on the games began in July 2014 and the MOBOT and Tiltfactor teams held bi-weekly meetings in which we made foundational decisions on: 1) target audiences and devices; 2) how to handle data with special characters and human-transcribed data; and 3) data import and export formats. By September, Tiltfactor had iteratively developed and playtested rough prototypes of two games 1) Smorball –for those with gaming experience (appendix 2) and 2) Beanstalk - for those with little game experience who are more altruistically motivated (appendix 3). In October, Tiltfactor developed the project's Backend Structure which holds the core database and data storage functionality. Further testing of game inputs identified some data quality issues which MOBOT is currently addressing (see Findings/Accomplishments section)

Objectives 6 & 7 : Evaluating accuracy scores from the game against ground truth pages and Generating error matrix for cleanup

Activity on objectives 6 & 7 will begin in September 2015.

Objective 8: Producing report and disseminating findings

We share progress on the project through a variety of in person presentations, traditional media and social media outlets.

Presentations

- TDWG Annual Meeting, Jonkoping, Sweden, Oct 2014, "Making Links in the BHL: Primary Source Materials as a Window to a Scientists' Methods" , Connie Rinaldo
- Council on Botanical & Horticultural Libraries meeting, Richmond VA, May 2014, "Purposeful Gaming, OCR correction and Seed & Nursery Catalog Digitization" , Marty Schlabach
- iDigBio CITScribe Hackathon, Florida, Dec 2013 "Purposeful gaming and BHL: engaging the public in improving and enhancing access to digital texts" , William Ulate

Media Coverage

- Harvard Gazette article about William Brewster's ornithological writings "Crowdsourcing old journals"
- BHL blog post on seed catalog digitization "The Stories Seeds Tell"
- Ornithology Exchange article about crowdsourcing transcription of William Brewster's ornithological writing "Step back into ornithological history"
- BHL blog post on the crowdsourcing aspects of the project "Crowdsourcing and BHL"
- BHL blog post on transcription activities "Transcribing the Field Notes of William Brewster"

- MOBOT blog post announcing choice of game designer, Tiltfactor http://www.missouribotanicalgarden.org/media/news-releases/article/700/missouri-botanical-garden-project-selects-designer-for-purposeful-gaming-grant.aspx
- BHL blog post announcing choice of Tiltfactor http://blog.biodiversitylibrary.org/2014/06/game-laboratory-tiltfactor-selected-for.html

- IMLS announcement http://www.imls.gov/news/2013_ols_grant_announcement.aspx#MO
- MOBOT press release http://www.missouribotanicalgarden.org/media/news-releases/article/639/garden-aims-to-improve-access-to-digital-texts-through-online-gaming.aspx

- Front page article in St Louis Post Dispatch http://www.stltoday.com/news/local/metro/missouri-botanical-garden-builds-a-new-kind-of-video-game/article_caa6455e-a789-58de-bf5e-9ba27f1c7856.html.
- Harvard Library http://lib.harvard.edu/blog-post-topics/ernst-mayr-mc
- Cornell http://mannlib.cornell.edu/news/new-games-old-seed-catalogs and http://news.cornell.edu/essentials/2013/12/games-purpose
- D-Lib magazine write-up in the In Brief column http://www.dlib.org/dlib/january14/01inbrief.html

Changes: none

Findings/Accomplishments:

Throughout the first year the project team has encountered several challenges and here we detail the methods we have used for overcoming them.

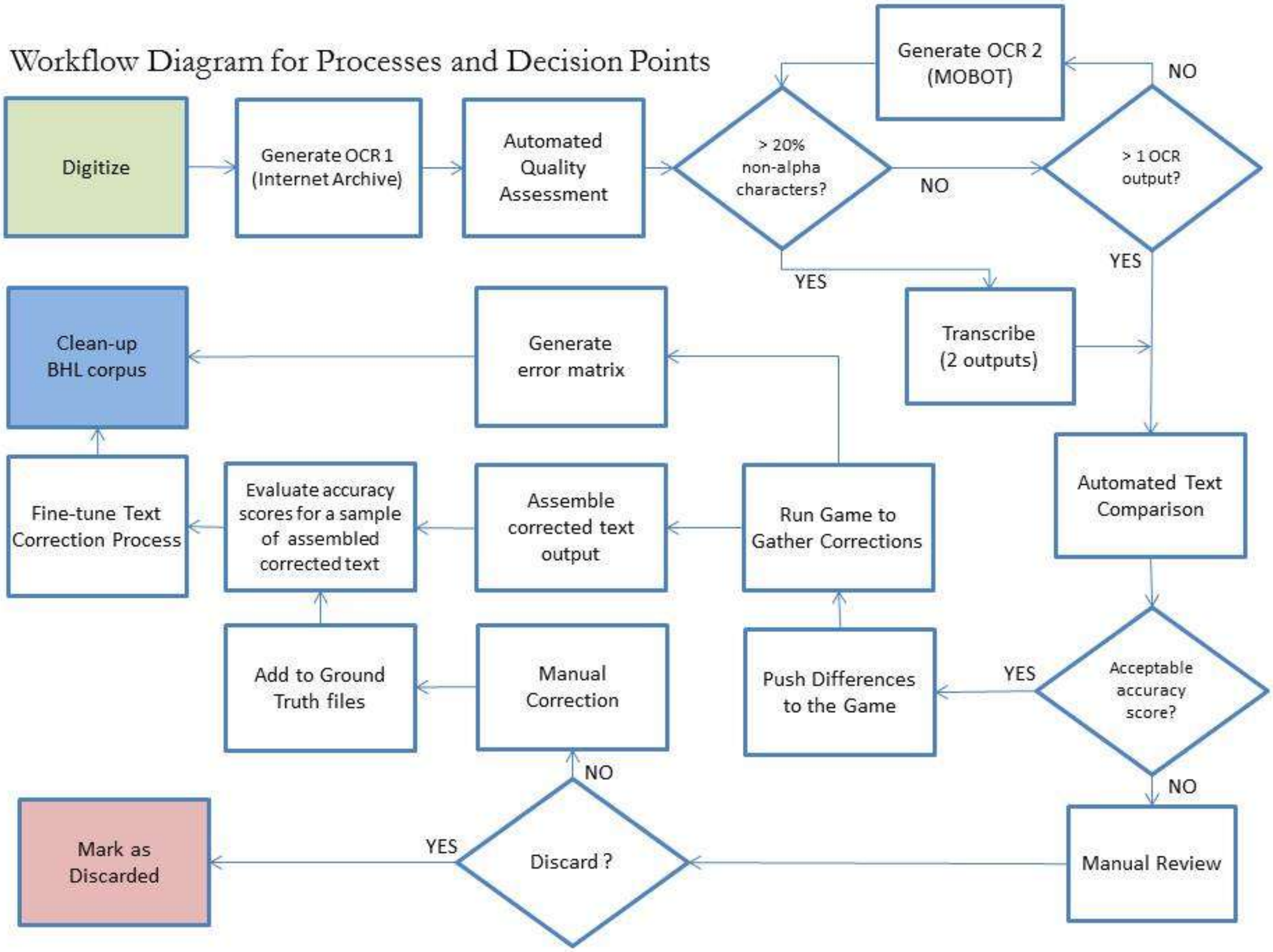
Minimize duplication of digitization efforts – In order to avoid digitizing the same content, project partners compared their holdings records. Historically, each of the institutions has taken different approaches to cataloging and describing seed and nursery catalogs, making comparison of records extremely difficult. To address this, staff at Cornell used an open source tool called Open Refine (http://openrefine.org/) to normalize the data to identify possible matches. This allows more accurate comparison of holdings which helps to further minimize duplication of effort. Equally important was the inclusion of NAL records during this process. Benefits included: a reduced duplication of effort, an increase in access to catalogs from NAL by bringing them into the collection within BHL, and more effective utilization of grant money. In addition, NAL became a BHL affiliate this fall which will help BHL strengthen its agricultural-related content.

Transcription coordinates – as mentioned above coordinate data is needed to be able to link the text output of a word to an image of the word on a page. OCR software automatically generates this data but transcription tools do not. None of the seven transcription tools that we reviewed could provide this functionality. Tiltfactor had proposed developing a transcription tool for us in lieu of a second game but we decided it was not the best use of the $110 dollars that was designated for game design. We also became aware of a tool being developed in Australia called TILT2 (https://github.com/AustESE-Infrastructure/TILT) that could provide this functionality. We have exchanged several data samples with the developer and are hopeful this tool will be finalized in time for us to use it on the project.

OCR output data quality issues – while testing inputs to the gaming platform we became aware of a problem in quality with the second OCR we were generating. In some cases, the OCR could not interpret some words and therefore some input words were left blank and coordinate data was often incorrect in these cases. We are currently investigating more current versions of the OCR software and how settings affect the outputs accuracy and are seeing more positive results.

Appendices

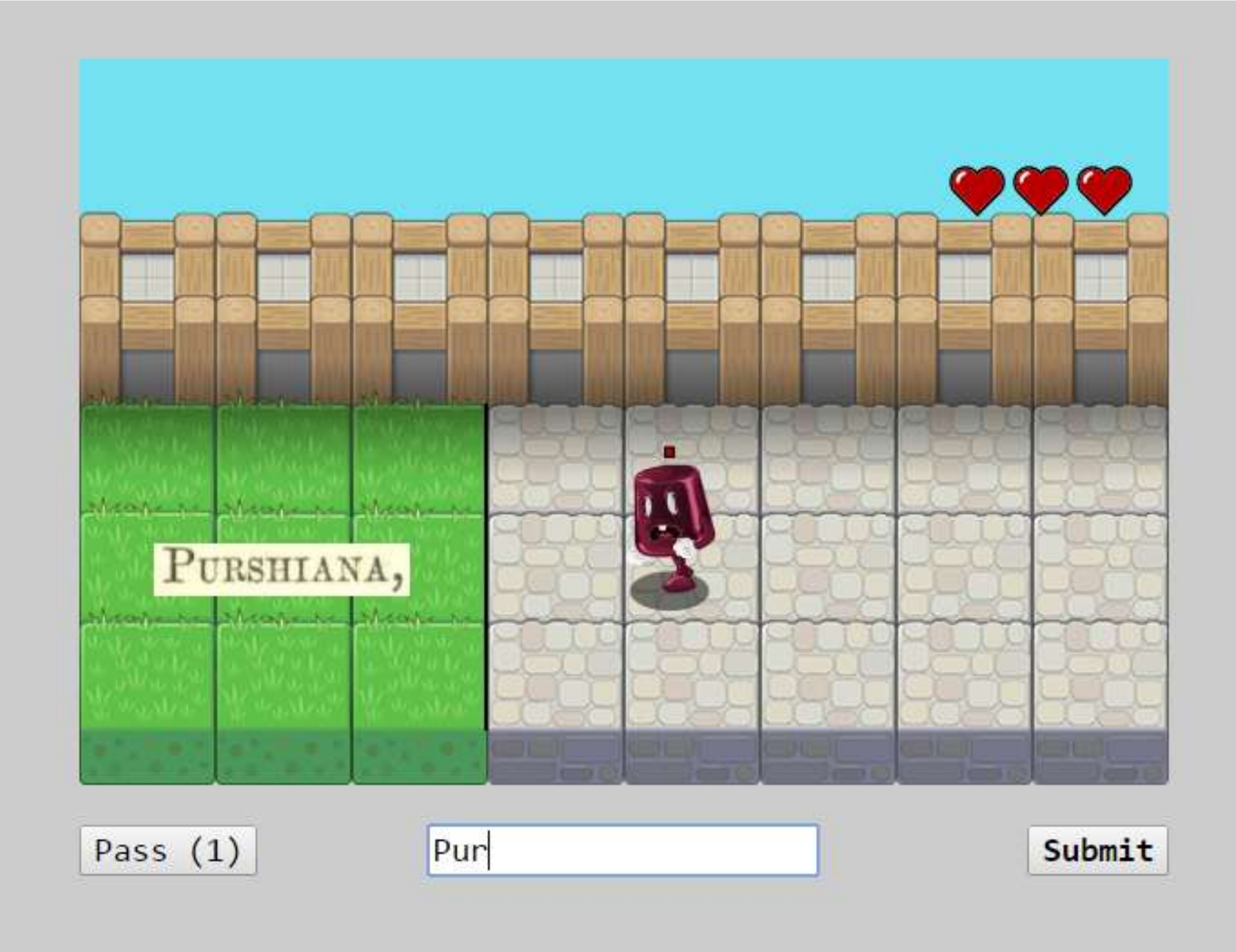
Appendix 1 Workflow Diagram


Workflow Diagram for Processes and Decision Points
Digitize
Generate OCR 1 (Internet Archive)
Automated Quality Assessment
> 20% non-alpha characters?
NO
YES
Generate OCR 2 (MOBOT)
> 1 OCR output?
NO
YES
Transcribe (2 outputs)
Automated Text Comparison
Acceptable accuracy score?
YES
NO
Push Differences to the Game
Run Game to Gather Corrections
Generate error matrix
Clean-up BHL corpus
Assemble corrected text output
Evaluate accuracy scores for a sample of assembled corrected text
Fine-tune Text Correction Process
Manual Review
Discard ?
YES
NO
Mark as Discarded
Manual Correction
Add to Ground Truth files

Appendix 2 Screen shot of current game prototype for gamers - Smorball

In *Smörball* you play the coach of the Harrisburg Melonballers, a fictional team that competes in the fictional and goofy amalgam sport of Smörball. Through shrewd tactics and excellent typing ability, you must coach, command, and direct your players to victory! Your team plays defense in every game of Smörball, and must prevent the opposing players from getting across the field to score in your endzone. As the coach, you must shout (type) commands to your players to tell them when to tackle their opponents. The fewer points the opposing team scores in each match, the better for you (and your accumulated winnings)!

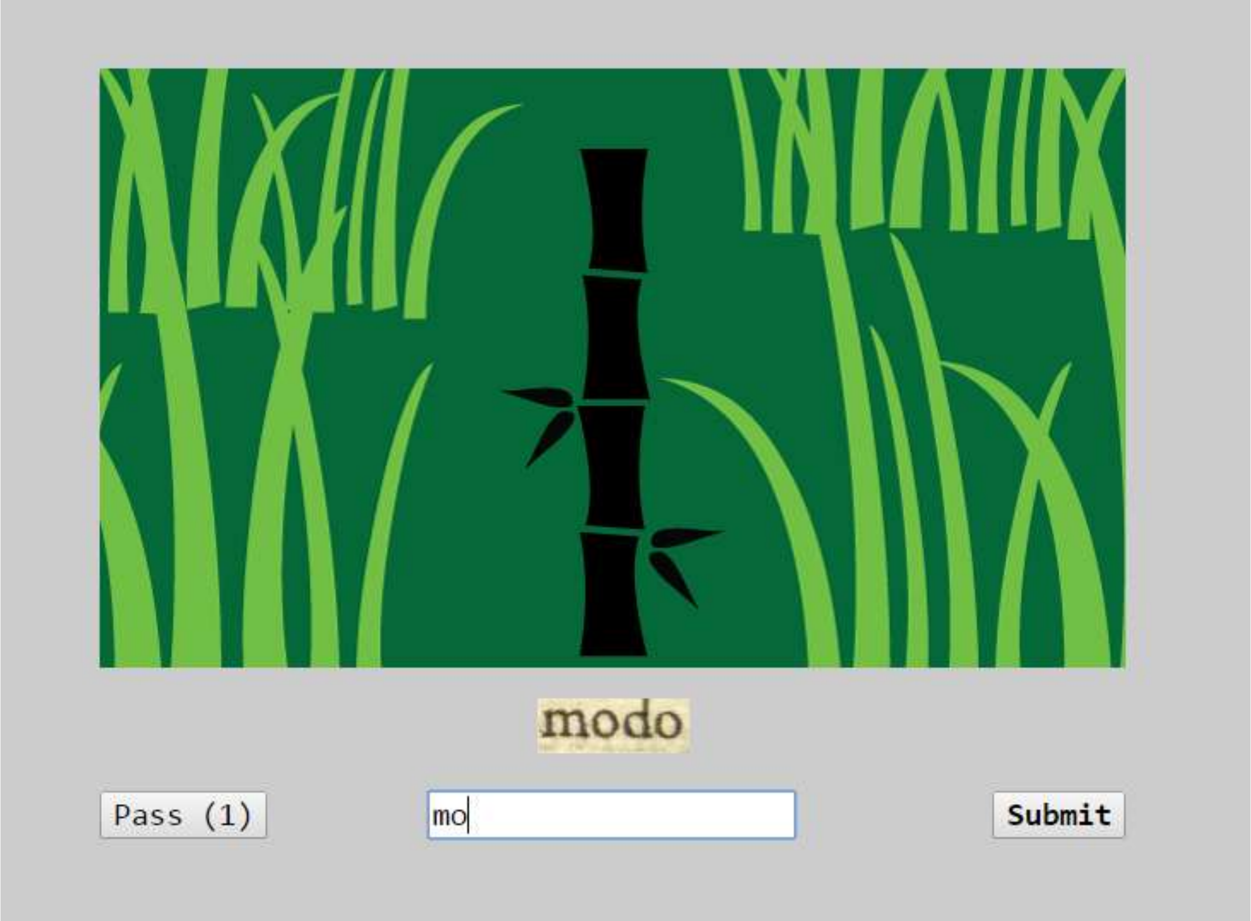
Note: This demonstrates the basic functionality and mechanics of the game but does not represent the final graphics



Appendix 3 Screen shot of current game prototype for non-gamers Beanstalk

In *Beanstalk*, players tend a small plant that must be spoken to in order to grow. As the player types words correctly, leaves and flowers sprout from the beanstalk. After 8 leaves have sprouted, the stalk grows a new segment with no leaves, and the count resets. Should the player make a mistake while typing a word, all the leaves they have typed so far on this segment wither and fall off. The plant starts off as a seedling, growing among blades of grass, and over the course of the game grows past bushes, trees, skyscrapers, clouds, and eventually becomes a massive beanstalk swaying among the stars. When the player reaches the stars her beanstalk drops another seed, and she can start a second stalk.

Note: This demonstrates the basic functionality and mechanics of the game but does not represent the final graphics


modo
Pass (1)
mo
Submit