Gamification of Private Digital Data Archive Management

Carlos Maltzahn Arnav Jhala Michael Mateas Jim Whitehead
University of California, Santa Cruz
{carlosm,jhala,michaelm,ejw}@soe.ucsc.edu

ABSTRACT

The super-exponential growth of digital data world-wide is matched by personal digital archives containing songs, ebooks, audio books, photos, movies, textual documents, and documents of other media types. For many types of media it is usually a lot easier to add items than to keep archives from falling into disarray and incurring data loss. The overhead of maintaining these personal archives frequently surpasses the time and patience their owners are willing to dedicate to this important task. The promise of gamification in this context is to significantly extend the willingness to maintain personal archives by enhancing the experience of personal archive management.

In this paper we focus on a subcategory of personal archives which we call private archives. These are archives that for a variety of reasons the owner does not want to make available online and which consequently limits archive maintenance to an *individual activity* and does not allow any form of crowdsourcing out of fear for unwanted information leaks. As an example of private digital archive maintenance gamification we describe InfoGarden, a casual game that turns document tagging into an individual activity of (metaphorically) weeding a garden and protecting plants from gophers and includes a reward system that encourages orthogonal tag usage. The paper concludes with lessons learned and summarizes remaining challenges.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Indexing methods; H.3.7 [**Digital Libraries**]: User issues; K.8.0 [**General**]: Games

Keywords

Tagging, Casual games, Digital data archives

1. INTRODUCTION

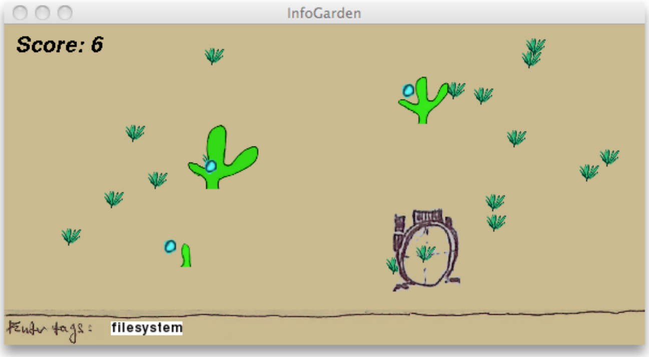
Digital data in personal and institutional environments is growing exponentially and faster than expected [10]. Personal digital archives such as email, photos, music, and movies already dwarf collections of physical documents in terms of their number of items. As opposed to physical collections, digital collections will unlikely ever run out of space given the sinking cost and ever increasing densities of digital storage media. But physical documents can survive tens, hundreds or thousands of years of inattention while digital documents require continual maintenance for their preservation [5, 3]. The combination of easy storage and difficult maintenance leads to a cycle of ever greater data amassment followed by ever greater amount of data loss. This is confirmed by Cathy Marshall's studies of digital archiving behavior of individual users. She sums up her findings with "It's easier to keep than to cull but it's easier to lose than maintain" [15]. Consequently digital data appears to be ephemeral and have lead to predictions that the current era would be known as the Digital Dark Age [1] leaving future historians deprived of any personal letters, diaries, and photo collections, sources that traditionally have proven essential for historic understanding.

All approaches to digital archive management known to the authors require the effort and discipline comparable to one or more full-time jobs. Most users of personal collections of photos, music, or movies seem to be unable to devote the attention to their personal archives to significantly increase the likelihood of archive survival. A recent effort that is partially addressing this situation is Facebook's Timeline: it provides an easy format to embed personal media such as photos and movies into rich contexts. Timeline is motivated by sharing via Facebook and reduces individual effort by leveraging various forms of crowdsourcing.

But what about collections that an owner for various reasons does not want to make available online? We call these *private archives* and note that their maintenance fundamentally depends on the individual efforts of the archive's owner. Examples of private archives are family photo and movie collections where the owner wants to protect the privacy of family members, and collections of scientific publications where the owner does not want to share personal annotations, taxonomies, and associations that are part of research yet to be published.

In this paper we are investigating the gamification of private archive maintenance – maintenance which lacks the motivation of sharing and the leverage of crowdsourcing. We are using a casual game approach and hypothesize that *users are able and willing to spend significant time and effort on archive management if that activity is re-casted as a fun and entertaining casual, single-user computer game*. Such a game could tap into the significant pool of "solitaire cycles" that users usually spend on playing traditional unproductive and distracting casual games. According to the 2007 report by the Casual Games Association, casual gamers average 7-15 hours of (online) play a week [6]. We chose to focus on *casual* game designs such as Solitaire or Tetris because of their simple rules and the ability to play them without much commit-

GamifIR '14, April 13 2014, Amsterdam, Netherlands



Enter tags: filesystem

Figure 1: The Garden Patch. Weeds and plants with fruit are placed in a garden patch. The player targets a weed, plant, or fruit with a reticle and enters keywords into the field at the bottom. By hitting the Return key the keywords are "shot" at the target. ©Carlos Maltzahn

ment so they fit within short work breaks. More people play casual games than any other type of video game [6].

The paper first gives an overview of InfoGarden (presented as a Work-in-Progress and Poster [14]) and then describes lessons learned and summarizes remaining challenges.

2. RELATED WORK

Gamification has garnered a lot of interest in recent years [9, 16] (see also a recent survey in [11]). One of the first examples known to the authors is Dennis Chao's PSDOOM [7] in which the administrative management of processes is turned into a first-person shooter game similar to Doom (by Id Software, Mesquite TX, 1993). Luis van Ahn's ESP Game [18] turns a tedious and expensive task of image labeling into a game that not only is successfully garnering free labeling resources but also provides verification of its results [12]. Another very successful example is *Fold.it* and its role in getting the public involved in science through elements of gamification [8]. Recently Defense Advanced Research Projects Agency (DARPA) as part of its Crowd Sourced Formal Verification (CSFV) program released five games at a verigames.com that gamify crowd-sourced formal verification of software. One of the authors of this paper is involved in one of the games, Xylem [17].

However, with the exception of PSDOOM, all above examples heavily rely on crowdsourcing. A successful example of gamification without crowdsourcing are the games DragonBox which teaches algebra [13] and memrise which teaches languages [2]. The latter also uses a gardening metaphor: completing levels of the game turns seeds into plants.

3. INFOGARDEN

We designed and built an initial prototype, InfoGarden, that focusses on a common and particularly tedious aspect of archive management: tagging. InfoGarden uses the metaphor of gardening and represents the status of a digital archive as a garden patch. Neglect of the archive is represented by the spread of weeds (see Figure 1).

The player can tend to the garden patch by weeding. The game is casual as it can be interrupted at any time and its elements and rules are very simple: the elements of the game consists of weeds, plants, fruit, gophers, and a reticle (Figure 2).

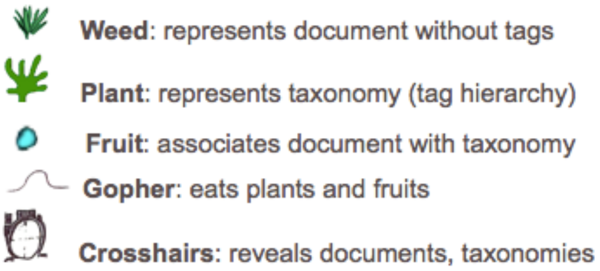

Weed: represents document without tags
Plant: represents taxonomy (tag hierarchy)
Fruit: associates document with taxonomy
Gopher: eats plants and fruits
Crosshairs: reveals documents, taxonomies

Figure 2: The Game Elements. The elements of the game are three pieces of vegetation (weeds, plants, and fruit), gophers, and a reticle. We carefully chose the semantics of plants and fruit based on observations of tag structures and effective tagging practices. ©Carlos Maltzahn

Documents without keyword annotations crop up as weeds at random places on the garden patch. Targeting a weed with the reticle will provide a view of the document it represents (Figure 3). Projecting a keyword at the weed removes the weed and a new fruit will appear at a plant. Plants represent hierarchies of tags with proper subset relationships in terms of tagged documents, also known as taxonomies: a child keyword is associated to a proper subset of the documents associated with the parent keyword.

A fruit represent a new association of a document with one or more of the keywords represented by the plant. A weed can be tagged with multiple keywords which can result in multiple fruit at multiple plants depending on how semantically independent the keywords are (Figure 4), i.e. how few subset relationships exist among these keywords. Keywords can also be projected at existing fruit to update tagging of documents.

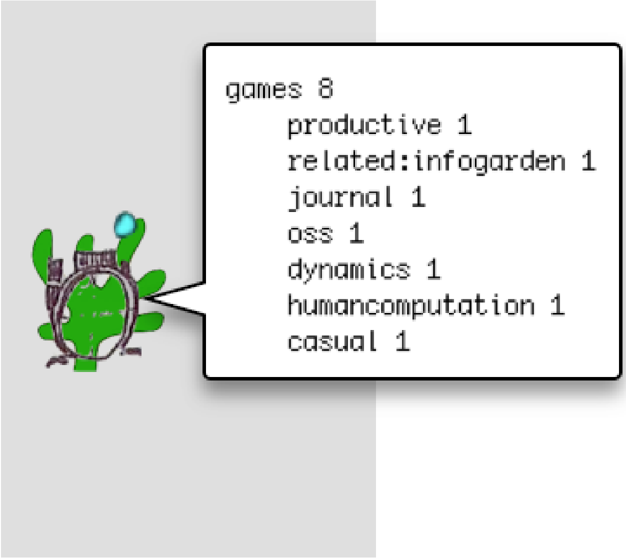
Gophers add elements of urgency, strategy, and levels of difficulty to the game. Gophers hide behind weeds and eat plants. The game is divided into rounds of increasing difficulty. The player wins a round once all weeds have been removed and looses when gophers managed to eat all plants before all weeds were shot. The difficulty is increased by the number of gophers initially roaming the garden patch.

3.1 First-Person Shooter Elements

The game borrows elements from a first-person shooter design by representing the tagging of a document as targeting a weed in a reticle, shooting keywords at the targeted weed with the sound of a gun, and letting the weed disappear in a fiery explosion with a satisfying boom (Figure 5). Shooting keywords can also kill gophers that happen to hide behind the targeted weed. Gophers move between weeds at random times and eat plants they come across on the way. They prefer to move to weeds and eat plants that are nearby. By clearing weeds near plants first the player creates greater distances between plants and weeds and reduces the likelihood that a gopher eats a plant. However, once enough weeds are removed so that no more nearby weeds exist and the ratio of plants over weeds has increased, gophers will consider longer distances and more likely eat plants. Thus, to win the game the player needs to focus on weeds with gophers and shoot keywords at them before gophers move on.

3.2 Scoring Points

Scoring points is one of the most common design elements of game play and allows different players to use the game to compete for higher scores. In InfoGarden shooting weeds, gophers, and fruit scores points. Scoring is designed to encourage "orthogonal" tag-



```
games 8
    productive 1
    related:infogarden 1
    journal 1
    oss 1
    dynamics 1
    humancomputation 1
    casual 1
```

Figure 3: Revealing Taxonomy. Hovering the reticle over a plant reveals the taxonomy it represents. In this case, "games" is used as keyword for eight documents and it has a proper subset relationship with each of the seven keywords underneath it, e.g. "dynamics" is used as keyword for one document together with "games". ©Carlos Maltzahn

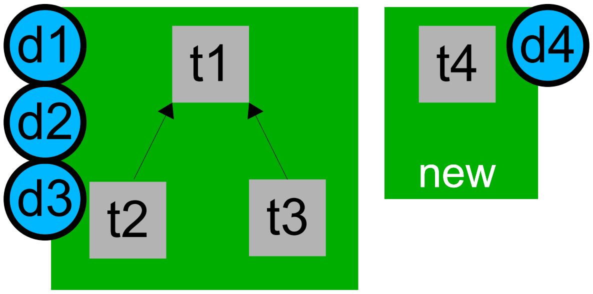


Figure 4: Planting. Tagging that removes a proper subset relationship between keywords or that introduces a new keyword without any subset relationship with other keywords creates a new plant. For example: given d1(t1), d2(t1, t2), d3(t1, t3) (where d1(t1) denotes a document d1 tagged with tag t1) then d4(t4) creates a new plant. Similarly, d5(t2) would create a new plant because it would remove the subset relationship between t1 and t2. ©Carlos Maltzahn

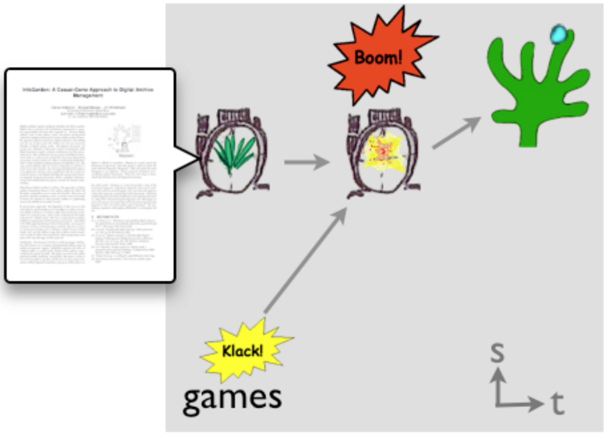


Figure 5: Shooting Weeds. Borrowing from first-person shooter game designs the player first targets a weed. Hovering the reticle over a weed reveals the document that the weed is representing. The player then types in one or more keywords and hits return. This triggers the sound of a gun and the keywords fly to the targeted weed. When they arrive they cause the weed to explode in a fire ball with the sound of an explosion. Fruit pop up on one or more plants with the score counter advancing accordingly with a mechanical clicking sound similar to a pinball machine. ©Carlos Maltzahn

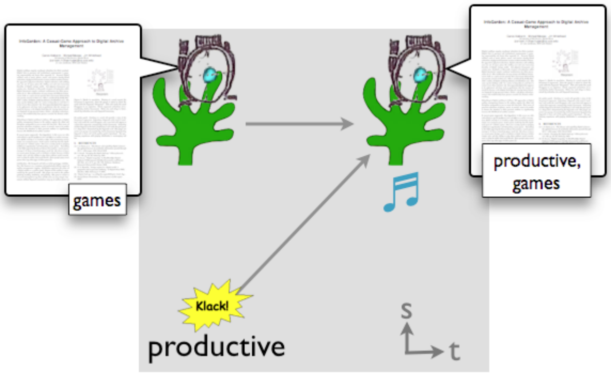


Figure 6: Shooting Fruit. Similar to shooting weeds, hoveringthe reticle over a fruit reveals the document it is representing and allowing its keywords to be edited. Hitting the return key sounds a gun shot and the edited keywords are flying towards the fruit. When they arrive, a bell sounds, the original fruit disappears, new fruit pop up at one or more plants, and the score advances. ©Carlos Maltzahn

ging: a shot weed scores a point plus one additional point for each newly created fruit. Keywords are considered orthogonal if there is no proper subset relationship between them. Thus, in the best case, each keyword creates a new fruit at a different plant. The intuition behind orthogonal tagging is that it covers more facets of the document and increases the likelihood of a user to recall the tagged content [19]. Shooting a weed also scores a point for each gopher hiding behind it at the time of explosion. Scoring for shot fruit is similar with the same incentive for orthogonal tagging. This encourages a strategy to quickly shoot weeds with single keywords (and kill gophers in the process) and then refine tagging of documents later, effectively partitioning each round into a fast-paced "hunting" phase and a more more contemplative "planting" and "fertilizing" by creating new plants and multiplying fruit via editing keywords of documents tagged during the hunting phase (Figure 6).

When gophers manage to eat plants the player looses points equal to the number of fruit the eaten plant was carrying. Note that this does not remove tagging from documents – the useful side effect of the effort of playing InfoGarden is preserved and as soon as the player tags a document with at least one of the keywords represented by the eaten plant, it will re-appear with all its fruit. But scoring will only include the count of the new fruit created by the most recent tagging. However, when a gopher manages to eat the plant again, all fruit carried by the plant will count against the player's score. Thus, large plants represent a risk of higher point losses thereby creating another incentive for players to choose many orthogonal keywords that create many small taxonomies instead of a few large ones.

3.3 Plant Morphology and Placement

Plants represent tag taxonomies based on proper subset relationships. In InfoGarden plants can represent two dimensions: the number of documents tagged with keywords included in the taxonomy is represented by the size of the plant using

$$size = \log_2 |taxonomy|_{docs}$$

while the number of keywords included in the taxonomy is represented by the branching of the plant,

$$branching = \log_2 |taxonomy|_{tags}$$

. InfoGarden uses 42 different plant shapes in seven sizes and with up to six branches to represent taxonomies containing up to $2^6 = 64$ tags and covering up to $2^7 = 128$ documents (Figure 7). Anecdotal evidence of personal collections seem to indicate that these scales are sufficient. Fruit is placed at a random location on the plant. Plants are planted on a random location on the garden patch.

We chose this very simple plant morphology and placement to focus on basic game dynamics. Obviously, plants and their placement can have much richer semantics. For example, the placement of plants could represent a topic map of tagged documents where the distance encodes the number of shared documents between two taxonomies. As another example, the placement of fruit could encode the position of the keyword in the taxonomy represented by the plant. We will return to this aspect of InfoGarden's design in Section 5.

3.4 Scalability

Today's personal computer typically stores 1-2 million files. Even when focussing on managing the collection of only one file type the number of untagged documents can be easily much larger than can be easily displayed on a garden patch (Figure 8). This poses two scalability challenges to InfoGarden: one in space where only a limited number of weeds can be displayed on a garden patch, and one in time where the task of tagging a large number of documents can be too exhausting and overwhelming and incompatible with the motivation for casual gaming. We addressed both of these challenges by using game rounds: each round begins with 20 weeds from randomly selected untagged documents and no plants. As tagging progresses, the associated plants show up in shapes that represent the taxonomies of already tagged documents. As the player successfully completes rounds, the difficulty is increased by introducing more gophers to keep the game interesting.

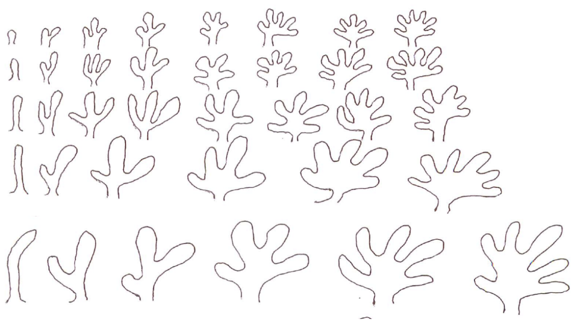


Figure 7: Plant Morphology. We drew highly simplified plant shapes with pencil on paper along two dimensions, size and number of branches. After scanning in this drawing, we partitioned each plant into separate images and named each image by its size and number of branches. We use 42 different plant shapes representing 7 sizes and up to 6 branches. Branching represents the number of the taxonomy's keywords and the size represents the number of documents tagged by any of these keywords. ©Carlos Maltzahn

4. LESSONS

Pygame as a prototyping platform. A paper and pencil and a scanner was all we needed to create the imagery for InfoGarden. It took only 1,500 lines of Python code to program the behavior of the imagery, making for a very fun and productive rapid prototyping experience.

The power of simple rules. We were surprised how simple rules can give rise to complex incentives and interactions. An example is the very simple scoring mechanism of counting newly created fruit which created an incentive for good tagging that was otherwise difficult to quantify. Another example are the very simple rules of how a gopher moves around: pick randomly a time within a time interval and pick randomly one of the five nearest weeds or plants as the next target. This gives rise to the strategy to clear weeds around plants to keep gophers away from them. It also provides incentive to focus on shooting weeds that have gophers hiding behind them.

Some people really disliked gophers, others loved them. Gophers are the most debated feature of InfoGarden. Many people we asked to evaluate the game would have preferred a more contemplative style and found that gophers were unnecessary and made the game unnecessarily hectic.

The game does not scale. One of the weakest points of InfoGarden is that it is still very tedious to tag documents. To have a chance to catch up with the daily influx of new documents, much more efficient forms of tagging such as simultaneous tagging of large selections of documents must be available. We are experimenting

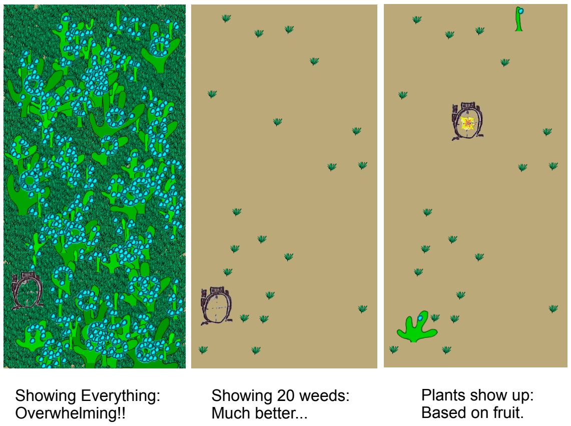


Figure 8: Scalability. The left-most garden patch shows all tagged and untagged pdf documents on the laptop of one of the authors. A much more manageable amount of information is displayed in the middle patch which shows a typical configuration at the beginning of a round: InfoGarden randomly selected 20 untagged documents and hides all existing taxonomies to minimize clutter. The right patch shows a patch after the round has progressed, revealing existing taxonomies (or creating new ones) when one of their keywords is used for tagging. ©Carlos Maltzahn

with using automatic pre-analysis of untagged documents to partition them into semantically similar clusters and representing these clusters with distinct representations of weeds.

5. REMAINING CHALLENGES AND CONCLUSIONS

The key challenge of InfoGarden, and any other private archive maintenance game is to make tagging fun and efficient even in large document collections. An instructive example might be Apple Computers' 2009 introduction of face detection in iPhoto: the photo library detects location of faces in photos and creates a profile of each. Users can now simply fill in names into faces and the library will then suggest faces it considers similar to those already filled in. Smart use of auto-completion and other user-interface components reduces this task to a few key strokes and makes it very fast. Similarly, a future version of InfoGarden will automatically analyze the entire collection of documents using, say, probabilistic topic modeling [4] (assuming textual documents) and then use these topic models to identify sets of weeds that can be tagged with the same set of keywords. The result of automatic analysis could also be used for level design: weeds occurring in a particular round could be limited to a small number of topics which would provide more opportunity for tagging entire sets of weeds.

Most of the focus of InfoGarden so far is on the activity of tagging documents. Another important part of archive management is the maintenance of good taxonomies. Over time the meaning of keywords might drift and new documents might require new keywords. In both cases the design of the keyword taxonomies would have to be revisited. And any change would require all affected documents to be updated. So it is not sufficient to just incentivize good design by scoring. Taxonomy management in InfoGarden will require a more sophisticated representation of taxonomies as plants, using their placement in the garden patch and the placement of fruit on plants. Following the gardening metaphor, players should be able to directly manipulate the placement of plants and their taxonomies.

In conclusion, we found that gamification of private archive management is very promising but will need additional information retrieval techniques to be effective and efficient.

6. REFERENCES

[1] Digital dark age. en.wikipedia.org/wiki/Digital_Dark_Age.

[2] memrise. Web Page. www.memrise.com, 2014.

[3] H. Besser. Digital longevity. In M. Sitts, editor, *Handbook for Digital Projects: A Management Tool for Preservation and Access*, chapter 9, pages 164–176. Andover: Northeast Document Conservation Center, 2000.

[4] D. M. Blei. Probabilistic topic models. *CACM*, 55(4):77–84, 2012.

[5] S. Brand. Escaping the digital dark age. *Library Journal*, 124(2):46–49, February 1999.

[6] Casual Games Association. Casual games market report 2007.

[7] D. Chao. Doom as an interface for process management. In *CHI '01*, pages 152–157, New York, NY, USA, 2001. ACM.

[8] S. Cooper, F. Khatib, and D. Baker. Increasing public involvement in structural biology. *Structure*, 21(9):1482–1484, 3 September 2013.

[9] S. Deterding, D. Dixon, R. Khaled, and L. Nacke. From game design elements to gamefulness: Defining "gamification". In *MindTrek'11*, Tampere, Finland, September 28-30 2011.

[10] J. F. Gantz et al. The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011. White paper, IDC, March 2008.

[11] J. Hamari, J. Koivisto, and H. Sarsa. Does gamification work? — a literature review of empirical studies on gamification. In *47th Hawaii International Conference on System Sciences*, Hawaii, USA, January 6-9 2014.

[12] E. Law and L. von Ahn. Input-agreement: A new mechanism for collecting data using human computation games. In *CHI 2009*, Boston, MA, April 4-9 2009.

[13] J. Liu. Dragonbox: Algebra beats angry birds. *Wired, June*, 2012.

[14] C. Maltzahn, M. Mateas, and J. Whitehead. Infogarden: A casual-game approach to digital archive management. In *Work-in-Progress and Poster Session at FAST'10*, San Jose, CA, February 24-27 2010.

[15] C. C. Marshall. Benign neglect in a digital world: a pragmatic look at personal archiving. In *Digital Lives 2009*, London, UK, February 9-11 2009.

[16] S. Nicholson. A user-centered theoretical framework for meaningful gamification. In *Games+Learning+Society 8.0*, Madison, WI, 2012.

[17] T. Stephens. UCSC computer scientists turn software verification into gameplay. Web Page. news.ucsc.edu/2013/12/xylem-game.html, December 6 2013.

[18] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI 2004*, Vienna, Austria, April 24-29 2004.

[19] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. In *Proceedings of Collaborative Web Tagging Workshop at 15th International World Wide Web Conference*. Citeseer, 2006.