Crowdsourced Knowledge Acquisition: Towards Hybrid-Genre Workflows

Marta Sabou, Arno Scharl
Department of New Media Technology
MODUL University Vienna
{marta.sabou, arno.scharl}@modul.ac.at

Michael Föls
Research Institute for Computational Methods
Vienna University of Economics and Business
michael.foels@wu.ac.at

ABSTRACT

Novel social media collaboration platforms, such as games with a purpose and mechanised labour marketplaces, are increasingly used for enlisting large populations of non-experts in crowdsourced knowledge acquisition processes. *Climate Quiz* uses this paradigm for acquiring environmental domain knowledge from non-experts. The game's usage statistics and the quality of the produced data show that *Climate Quiz* has managed to attract a large number of players but noisy input data and task complexity led to low player engagement and suboptimal task throughput and data quality. To address these limitations, we propose embedding the game into a *hybrid-genre workflow,* which supplements the game with a set of tasks outsourced to micro-workers, thus leveraging the complementary nature of games with a purpose and mechanised labour platforms. Experimental evaluations suggest that such workflows are feasible and have positive effects on the game's enjoyment level and the quality of its output.

Keywords: knowledge acquisition, crowdsourcing, games with a purpose, mechanised labour, CrowdFlower, workflow, climate change

1. INTRODUCTION

The difficulty of acquiring, representing and maintaining an intelligent system's knowledge base has been coined as the *knowledge acquisition bottleneck* in Artificial Intelligence (AI) research (Feigenbaum, 1977). More than 30 years later, this problem continues to affect not only the AI area but also the field of the Semantic Web where the goal of building an intelligent layer over the World Wide Web (Berners-Lee et al., 2001) is hampered by the lack of Web-scale knowledge resources, both in terms of domain models (i.e., ontologies) and instance annotations. Recent years have seen a tremendous increase of openly available formal knowledge resources on the Web thanks to the linked data movement (Heath & Bizer, 2011), in particular regarding information on the instance level. Terminological knowledge, however, is still scarce, especially in novel (or less popular) domains.

The social web enables new ways for collaborative knowledge creation, as a way to overcome the knowledge acquisition bottleneck of the Semantic Web. Social media platforms facilitate involving large and diverse populations of users in the knowledge acquisition process, in one of the following two ways. A first set of approaches piggyback on the data created as part of other

Web systems to derive useful knowledge assets. For example, folksonomy induction algorithms extract knowledge from folksonomies derived from social tagging systems such as Flickr (Stohmaier et al, 2012). Doan et al (2011) consider that such approaches use an *implicit crowdsourcing* strategy to acquire their data. In contrast, a second set of approaches subscribes to an *explicit crowdsourcing* strategy by building their own, dedicated application for acquiring knowledge through large-scale social participation. For example, traditional knowledge creation tools have been extended to enable collective knowledge creation, including the Protégé ontology editor (Tudorache et al, 2013) or the GATE linguistic annotation toolkit (Bontcheva et al, 2013). While these extensions primarily support the collaborative and distributed work of knowledge experts, an increasing trend is allowing large populations of non-experts to create knowledge through the use of novel social media collaboration platforms such as games or mechanised labour platforms.

Climate Quiz (apps.facebook.com/climate-quiz) is an example of such an approach: it is a game with a purpose deployed on Facebook that facilitates the creation of knowledge in the environmental domain by a large population of non-experts (Scharl et al, 2012). Our evaluation of the game detailed in Section 4 showed that, while it has attracted a high number of players, the heterogeneous domain relevance of its input data hampers player engagement, leads to short play times and affects the quality of the output. To overcome these limitations, we propose embedding the game into a *hybrid-genre workflow,* which splits the complex problem of knowledge acquisition into tasks performed by players as well as micro-workers. This workflow leverages the pros and cons of games and mechanised labour platforms to improve gaming experience and output data quality. Our experiments show that such workflows are indeed possible, although future work will further fine-tune the synchronisation and task management across the two genres. This paper makes the following contributions:

- Section 2 presents *a survey of knowledge acquisition through crowdsourcing*, concluding with recent trends in the field and a comparison of strengths and limitations of different crowdsourcing genres.
- *Climate Quiz Evaluation*. As an extension of the earlier presentation of this game in (Scharl et al, 2012), this paper provides an in-depth evaluation of the game including its comparison with evaluation details of previous games that target (linguistic) knowledge acquisition (Section 4).
- *Implementation and evaluation of hybrid-genre workflows.* We propose a novel concept to workflow integration, and exemplify its implementation with the Climate Quiz. We show experimentally that such hybrid-genre workflows are feasible and that they can improve results as compared to single-genre approaches (Section 5).

Section 6 concludes the paper with the main lessons learned, and an outlook on future work.

2. CROWD-BASED KNOWLEDGE ACQUISITION – AN OVERVIEW

In this section we provide background details about various crowdsourcing genres (Section 2.1) and outline how crowdsourcing is used to support factual knowledge acquisition (Section 2.2). In Section 2.3, we then summarize related work on crowdsourcing workflows, before Section 2.4 concludes with an overview of benefits and limitations of mechanised labour and games with a purpose as the two main crowdsourcing genres.

2. 1. Crowdsourcing

Crowdsourcing techniques allow outsourcing a task to *"an undefined, generally large group of people in the form of an open call"* (Howe, 2009). They are classified in a number of genres according to various dimensions, such as the motivation of human contributors (fun vs. altruism vs. payment), the way in which individual results are aggregated, and how quality is managed. The three key crowdsourcing genres according to Quinn and Bederson (2011) are:

Mechanised labour (MLab) is a type of paid-for crowdsourcing, where contributors choose to carry out micro-tasks and are paid a small amount of money in return (often referred to as micro-payments). The most popular platform for mechanised labour is Amazon's Mechanical Turk (MTurk) which allows requesters to post their micro-tasks in the form of Human Intelligence Tasks (or HITs) to a large population of micro-workers (often referred to as "turkers"). Most projects use crowdsourcing marketplaces such as MTurk and CrowdFlower (CF), where contributors are extrinsically motivated through economic incentives.

Games with a purpose (GWAPs) enable human contributors to carry out computation tasks as a side effect of playing online games (von Ahn & Dabbish, 2008). An example from the area of computational biology is the Phylo game (phylo.cs.mcgill.ca) that disguises the problem of multiple sequence alignment as a puzzle like game thus "intentionally decoupling the scientific problem from the game itself" (Kawrykow et al, 2008). The challenges in using GWAPs in scientific context are in designing appealing games and attracting a critical mass of players.

Altruistic crowdsourcing refers to cases where a task is carried out by a large number of volunteer contributors. To reduce the incentive to cheat (e.g., for money or glory), altruistic crowdsourcing approaches leverage the intrinsic motivation of a community interested in a domain. The Galaxy Zoo (www.galaxyzoo.org) project, for example, seeks volunteers with a latent desire to help with scientific research for classifying Hubble Space Telescope galaxy images. The project has attracted more than 250,000 volunteers which provided over 150 million galaxy classifications.

2. 2. Crowd-based Knowledge Acquisition

Knowledge acquisition has been addressed through games from as early as 2006 when von Ahn built Verbosity (von Ahn et al, 2006), a GWAP inspired from the Taboo game (where a narrator offers hints and a guesser must guess the concept), which collects a database of common-sense facts – see Table 1. Verbosity uses the cards metaphor to guide the types of hints that the narrator gives. For example, the "Type" card allows providing hints about the super-classes of the concept, by generating appropriate natural language templates to be filled by the narrator. This approach ensures that the game collects a broad range of relations, such as *type*, *purpose*, *is related*, *is opposite*, and a variety of spatial relations. Built one year after Verbosity, the Common Consensus GWAP focuses on acquiring a particular type of knowledge, namely goals (Lieberman, 2007). Inspired from the Family Feud TV show, the game asks players to answer questions such as "What are some things you would use to watch a movie?" in order to elicit common-sense knowledge about goals. The game relies on a handful of question templates which allow acquiring different types of knowledge about goals (e.g., parent and children goals or orthogonal connections between goals). As a multi-player game, players receive points in real time for all their answers that are also given by other players. Vickrey and colleagues (2008) report on three games (inspired from the Scattegories and Taboo games) that aim to collect semantically related words with their three games: Categorilla (players must supply a phrase fitting a specific category, e.g., “things that fly”, and starting with a given letter), Categodzilla (same as Categorilla but without the letter restriction) and Free Association, where players must type words related to a given “seed” word.

The emergence of Semantic Web research has reiterated the need for formally recorded domain knowledge. Early attempts of applying the GWAP paradigm to this area have primarily focused on alleviating human-input intensive tasks such as ontology learning and matching. For ontology learning, OntoPronto (Siorpaes & Hepp, 2008) is one of the games running on the OntoGame platform that aims to build domain ontologies from Wikipedia articles by mapping these articles to the most specific class in the Proton ontology. For ontology matching, SpotTheLink (Thaler et al, 2011) is a real-time quiz running on the OntoGame platform and aiming to align concepts from DBpedia (www.dbpedia.org) and the Proton upper level ontology (proton.semanticweb.org). A current trend is building games that make use of the large body of linked open data (LOD) to build new knowledge artefacts, provide useful applications, and try to improve LOD quality:

- GuessWhat?! (Markotschi & Völker, 2010) creates ontologies by exploring instance data available as linked open data. Given a seed concept (e.g., banana), the game engine collects relevant instances from DBpedia, Freebase and OpenCyc and extracts the main features of the concept (e.g., fruit, yellowish) which are then verified through the collective process of game playing.
- BetterRelation (Hees et al, 2011) is a two-player game that aims to detect correct and relevant statements about a topic: players need to decide which of two statements about an entity would have “come to mind first” and score points if their selections coincide.
- WhoKnows? (Waitelonis et al, 2011) and RISQ! (Wolf et al, 2011) are developed by the same group and have a similar mechanism: they use LOD facts to generate questions and use the answers to (1) evaluate property ranking, i.e. identify the most important/relevant property of an instance; (2) detect inconsistencies; (3) find doubtful facts. The obtained property rankings reflect the “wisdom of the crowd” and are an alternative to semantic rankings generated algorithmically based on statistical and linguistic techniques. The games, however, differ in the gaming paradigm they adopt. While WhoKnows?! uses a classroom paradigm and aims towards being an educational game, RISQ! is a Jeopardy-style quiz game. RISQ! is distributed both through Facebook and as a standalone application.
- UrbanMatch (Celino et al, 2012) relies on players’ mobility to link LOD concepts to representative images from an image database.

Mechanised labour has been used only in recent years. In 2010, Eckert and colleagues relied on MTurk micro-workers to provide concept relations in the philosophy domain. ZenCrowd (Demartini et al, 2012) uses a mixed human-machine workflow to solve the entity linking problem and shows that crowdsourcing can improve precision on average up to 14% over the best algorithmic matchers. CrowdMap (Sarasua et al., 2012) combines ontology matchers with crowdsourcing through a loosely coupled workflow. Finally, a mechanised labour version of the OntoPronto game has been implemented recently in order to compare the two crowdsourcing genres on a common task and data set (Thaler et al, 2012).

We conclude that the use of crowdsourcing is a popular approach in the knowledge acquisition field, with a predominant use of the GWAP genre. This is in contrast with other fields, in particular natural language processing (NLP) and databases, where mechanised labour approaches are more frequently employed than games (Wang et al, 2012). Two key trends in the knowledge acquisition domain are (1) an increased use of mechanised labour instead of games, and (2) the introduction of loosely-coupled workflows that augment algorithms with human input (Demartini et al, 2012; Sarasua et al, 2012). There is no work on combining different crowdsourcing genres within a single workflow, however, although such combination of tasks would be beneficial but currently prevented by a lack of understanding of the complementarities of these genres (Thaler et al, 2012). In the next subsection, we provide an overview of how crowdsourcing workflows are used for knowledge acquisition and for solving NLP tasks. We then conclude with a comparative discussion of how games and mechanised labour platforms complement each other, as a further reason for investigating hybrid-genre workflows.

Approach	**Genre**	**Type of knowledge**	**Workflow**
Verbosity *(von Ahn, 2006)*[1]	GWAP	Common-sense facts *e.g. milk is white*	None
Common Consensus *(Lieberman, 2007)*	GWAP	Common-sense goals *e.g. To write an email, use your PC*	None
Categorilla *(Vickrey, 2008)*	GWAP	Words that fit categories *e.g. Things that fly, Types of fish*	None
Free Association *(Vickrey, 2008)*	GWAP	Words related to a seed word *e.g. (submarine, underwater)*	None
OntoPronto *(Siorpaes, 2008)*	GWAP	Classification of Wikipedia articles to PROTON concepts	Two stage game
SpotTheLink *(Thaler, 2011)*	GWAP	Ontology concept mappings (Align DBpedia and PROTON)	Two stage game
GuessWhat?! *(Markotschi, 2010)*	GWAP	Complex concept definitions*; e.g.* *Banana: fruit & yellow & grows on trees*	Create - verify
BetterRelation *(Hees, 2011)*	GWAP	Importance ranks for LOD properties	None
WhoKnows? *(Waitelonis, 2011)*	GWAP	Importance ranks for LOD properties	None
RISQ! *(Wolf, 2011)*	GWAP	Importance ranks for LOD properties	None
UrbanMatch *(Celino, 2012)*	GWAP	Links between URLs and images.	None
InPhO *(Eckert, 2010)*	MLab (AMT)	Concept relations (philosophy domain) *e.g. Dualism sameAs Philisophy of mind*	None

[1] For brevity, references used in tables mention only the first author and the date of the publications.

ZenCrowd *(Demartini, 2012)*	MLab (AMT)	Links between entities in text and LOD *e.g. "Berlin" sameAs dbpedia.org/page/Berlin instances*	Machine and human computation
CrowdMap *(Sarasua, 2012)*	MLab (CF)	Concept mappings (equivalence, subsumption)	Machine and human computation
OntoPronto-MTurk *(Thaler, 2012)*	MLab (MTurk)	Classification of Wikipedia articles to PROTON concepts	None

Table 1: Overview of crowdsourcing based approaches for knowledge acquisition, including their genre, the produced knowledge type and the use of workflows

2. 3. Crowdsourcing Workflows

Crowdsourcing workflows define how multiple crowdsourcing tasks are combined together (or with external modules) and, as such, they offer an alternative to solving more complex problems that cannot be easily split into multiple simple, independent tasks, executed in parallel.

Already in 2005, Chklovski, a pioneer in the use of crowdsourcing for NLP, observed that contributors cannot only solve tasks but can also verify the work performed by their peers. This resulted in introducing the idea of "validating contributors" as an alternative quality assurance method to the classical inter-annotator agreement, which does not cover those paraphrases that are correct but only entered once (Chklovski, 2005). The PhraseDetectives game exemplifies this idea by being structured into two core tasks, one for detecting markables and a second one for verifying the originally provided annotations (Poesio et al, 2012). Such *create-verify workflows* have also been implemented on crowdsourcing marketplaces. For example, Callison-Burch (2009) routinely includes a second, verification HIT following a data creation HIT for tasks such as the acquisition of multiple reference translations. In the commercial area, CastingWords specializes in transcribing audio files using a workflow in which fragments of the file are first transcribed by workers, than a second set of workers evaluates and, if necessary, re-transcribes the first version of the transcriptions (Hoffmann, 2009). From the Semantic Web related approaches above, GuessWhat?! uses this mechanism by combining a guessing phase in which players assign concept labels to concept definition and a subsequent evaluation phase where the created concept-definition assignments are scored as correct or not.

For solving more complex problems, some researchers have developed complex workflows consisting of the combination of three or more HITs. For example, Negri and colleagues (2011) solve an entailment corpora acquisition problem by applying a "divide and conquer" approach and splitting this complex task into a pipeline of 5 simpler tasks, some of which verify the quality of the previous tasks (e.g., a task of modifying English sentences while still preserving their information is followed by a task that verifies whether the generated sentence is grammatically correct). Soylent, a word processor interface that allows crowdsourcing editing tasks such as shortening, proofreading and general editing of document snippets, relies on a "Find-Fix-Verify" workflow (Bernstein et al., 2010) which involves 3 steps: (1) finding the snippets of text that require editing, (2) performing the actual editing work and (3) verifying the results of step 2 and selecting one best option (when a single value is required) or filtering out the poor options (when many options are favourable, e.g., rephrasing suggestions). In the knowledge acquisition area, both OntoGame and SpotTheLink involve two stage workflows for solving a more complex task. In the case of SpotTheLink, for example, ontology matching is broken down into i) agreeing on a related concept and ii) then deciding on the actual relation between two concepts.

Another type of workflows combines human computation with algorithms through typical patterns such as that of active learning. For example, in the NLP field, research has focused on active learning, e.g. for sentiment classification (Brew et al, 2010) and named entity annotation (Laws et al, 2011). These approaches leverage machine classifiers to predict which samples are the most informative (e.g., by measuring disagreement between multiple classifiers) to reduce the number of crowdsourced judgments. The integration of human input within the algorithmic computations reduces significantly the amount of human input required, thus making crowdsourcing even more cost- and time-effective. In the area of knowledge acquisition, human-machine workflows are less tightly coupled than active learning based approaches and are limited to algorithms working in tandem with crowds to solve problems. For example, ZenCrowd (Demartini et al, 2012) uses a probabilistic model to combine the results of algorithmic and crowd-based entity linking: entity linking is first performed by algorithmic matchers, then instances with low confidence are verified by the crowd, and finally all results, both from the algorithms and the crowd are merged through a probabilistic model to compute the final result set. CrowdMap (Sarasua, 2012) invokes crowdsourcing tasks to support an ontology matching algorithm.

Some efforts also focus on defining generic crowdsourcing workflow engines, geared mostly towards MTurk, and, in theory applicable to any task type. TurKit (Little et al., 2010) relies on an iterative task execution model where turkers iteratively improve and evaluate an artefact until the required quality is reached, thus formalizing and automating an iterative version of the create-verify workflows. TurKontrol (Dai et al., 2010) offers a decision-theoretic planner to optimise the TurKit workflows for the best quality/cost ratio. CrowdForge (Kittur et al., 2011) has a MapReduce style approach to decompose and solve complex tasks, where each task is solved by a series of partition-map-reduce steps. All these workflows have been designed for crowdsourcing marketplaces (i.e., they focus on single-genre crowdsourcing) and, to our knowledge, none of these mechanisms have yet been adopted by the NLP or knowledge acquisition communities.

2. 4. Pros and Cons of Crowd-based Genres

The question of how crowdsourcing genres compare to each other arises naturally. So far, however, there are no clear answers to this question. In fact, in the semantic web area, we are aware of a single effort of comparing games vs. mechanised labour platforms by evaluating them on the same task (Thaler et al, 2012). In the, NLP area, this question has been addressed using a survey based approach by (Wang et al, 2012) as well as by (Chamberlain et al, 2013) who consider the success and the limitations of games for language resource creation and compare those to characteristics of mechanised labour which they gather from the literature. In this section, we sum up the discussions of the previous papers, in order to support our design of the hybrid-genre workflow. Both genres have their pros and cons when comparing them along the key dimensions of any knowledge creation project: cost, speed and data quality, as discussed next and summarized in Table 2.

Costs. Projects based on mechanised labour have very low initial setup costs, since they reuse the platform's job web tools. They also allow performing tasks for very small amounts of money, however, since typically multiple judgments must be collected for each task for quality assurance purposes, the acquisition price for large resources can be prohibitive. In contrast, games tend to have high up-front costs to implement their user and management interfaces, but then allow gathering data virtually for free (Poesio, 2013; Thaler, 2012). Poesio and colleagues (2013) take a close look at the cost reductions enabled by crowdsourcing genres in the case of large resources, on the scale of 1M tokens. They estimate that, compared to the cost of expert-based annotation (estimated as $1.000.000), the cost of 1M annotated tokens could be indeed reduced to less than 50% by using MTurk (i.e., $380,000 - $430,000) and to around 20% of the expert-based approach's price (i.e., $217,927) when using GWAPs, such as their own PhraseDetectives game. Therefore, mechanised labour is more cost effective for quick and affordable acquisition of small-scale datasets, while GWAPs can make larger content creation projects more affordable, thanks to their very low ongoing maintenance costs.

Speed. Crowdsourcing projects use throughput (the amount of data created per human hour) to measure the speed of data creation. Chamberlain and colleagues (2013) report on throughputs of 450 and 648 for the two annotation GWAPS they describe, however, these speeds remain far behind the almost real-time completion of tasks on MTurk. Thaler et al (2012) have also shown that the time needed to run the same experiment with the OntoPronto game was double to that needed when using MTurk. Indeed, paid-for crowdsourcing has the advantage of a faster and more predictable completion time, since projects tap into an already existing labour pool and reusable web interfaces for task and worker management. In contrast, completion times of GWAPs are often slower and much less predictable and depend on the ability to recruit, retain, and motivate a large number of contributors.

Characteristic	**MLab**	**GWAP**	**References**
Cost			
Set-up Price & Time	Low (+)	High (-)	(Poesio, 2013), (Thaler, 2012)
Price per task	Low (-)	None (+)	(Poesio, 2013), (Thaler, 2012)
Speed			
Throughput	High (+)	Low (-)	(Chamberlain, 2013)
Throughput Predictability	High (+)	Low (-)	(Chamberlain, 2013)
Data Quality			
Maintaining Motivation	Easy (+)	Difficult (-)	(Thaler, 2012)
Incentive to cheat	High (-)	(Mostly) Low (+)	(Wang, 2012)
Task Complexity	Simple (-)	Complex (+)	(Chamberlain, 2013)
Importance of task interestingness	Low (+)	High (-)	(Wolf, 2011), (Thaler, 2012)
Worker diversity	Low (-)	High (+)	(Thaler, 2012), (Parent, 2011)
Other issues			
Ethical issues	Yes (-)	No (+)	(Fort, 2011), (Eckert, 2010)

Table 2: A comparison of advantages (+) and disadvantages (-) of crowdsourcing genres

Quality. There are various factors that influence the quality of data that can be obtained through crowdsourcing. In games, maintaining a motivated player base is very difficult and often requires choosing (even manually) only interesting tasks (e.g., ontologies in a domain of interest). Micro-workers, on the other hand, are motivated extrinsically by pay and will accept tasks independently of their level of interestingness thus being suitable for a broader range of projects. Chamberlain et al. (2013) observe, however, that micro-workers have difficulties in performing complex tasks such as the evaluation of summarisation systems, which might otherwise be feasible with a stable player population that can be trained on a particular task. The extrinsic motivation of micro-workers has however downsides, namely that they are more likely to cheat to obtain an economic benefit than players who play for fun (Wang et al, 2012). A final quality related issue is data bias. Statistics from MTurk (Fort et al, 2011) and GWAPs (Poesio et al, 2011) have shown that a small number of people carry out a large number of tasks (paid HITs or hours playing), which, if the aim is to have more diverse data, from different people, might bias the results. Compared to paid-for marketplaces, GWAPs promise superior results, not only due to their intrinsically motivated players but also by making better use of sporadic, explorer-type users. In fact, recent studies show that games may provide a larger variety of contributors and can reach more individuals than MTurk (Parent and Eskenazi, 2011). Similarly, Thaler et al (2012) found that their game reached out to a larger player base (270) than when recruiting micro-workers on MTurk (only 16).

Ethical and legal issues related to crowdsourcing are an increasingly hot topic. The use of mechanised labour (MTurk in particular) raises a number of worker right issues: low wages (below $2 per hour), lack of worker rights, and legal implications of using MTurk for longer-term projects (Fort et al, 2011).

We conclude that there is a high complementarity among the game-based and mechanised labour crowdsourcing genres along all key dimensions (cost, speed, quality) and that this fact could be leveraged for building hybrid workflows. For example, complex, interesting tasks could be performed by a dedicated player base (on a longer term and virtually for free), while more routine (and therefore boring) tasks that would reduce the motivation of players might be more suitable for execution by extrinsically motivated micro-workers, for a small amount of money. We describe and evaluate such a workflow in Section 5.

3. CLIMATE QUIZ

Climate Quiz acquires knowledge in a specific domain (i.e., climate change), as opposed to harvesting generic knowledge as most of the current knowledge acquisition games do. As such, it appeals to environmental enthusiasts and leverages their interest in the domain as an additional motivational factor (besides the fun factor), thus being the first game in the knowledge acquisition area that includes some elements of altruistic crowdsourcing

The goals of Climate Quiz are twofold. Firstly, Climate Quiz acts as a “game with a purpose” with the main aim of collecting knowledge assets to support an ontology learning algorithm (Wohlgenannt et al., 2012). A human-machine workflow is therefore established as depicted in Figure 1. The “machine” part of the workflow is the ontology learning algorithm that extracts terms from unstructured and structured data sources. The term pairs that are most likely related based on the algorithm’s input data sources are subsequently sent to Climate Quiz, where the human element of the workflow assigns relations to these pairs. These relations are fed back into the algorithm which uses them to perfect the learned ontology and to derive new term pairs that should be connected. The ontology learning algorithm (Liu et al, 2005; Weichselbraun et al,

2010) has been continually refined over several years as part of the webLyzard text mining framework (www.webLyzard.com). It incorporates a range of methods from statistics, artificial intelligence and natural language processing, including co-occurrence analysis, subsumption analysis, link type detection, Hearst patterns, and spreading activation.

Secondly, Climate Quiz has a pronounced educational goal by aiming to raise awareness of climate change related issues. Players learn about these issues in the process of assigning relations to terms in this domain. The game playing process entices them to look up external sources to acquire the necessary knowledge for providing the right relation (e.g., not all players might know what "climate forcing" or "albedo" means but these terms can be easily looked up on the Web). Additionally, the game alternates relation assignment tasks with quiz-like questions about climate change, both as a mechanism to reduce the routine of assigning relations and as a way to provide further educational value. This educational aspect is a differentiating feature of Climate Quiz compared to other knowledge acquisition games, which makes the game resemble a virtual citizen science project (Wiggins & Crowston, 2011).

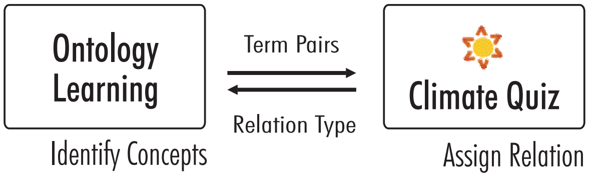
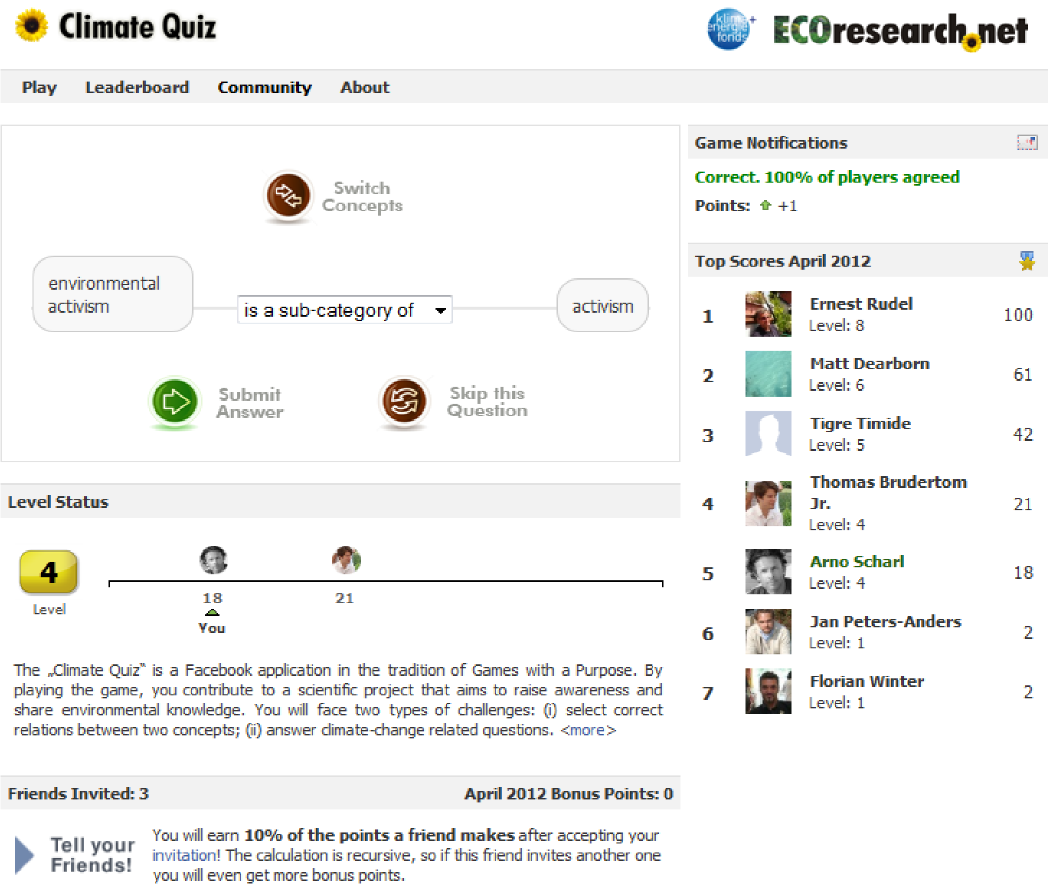


Figure 1: Human-machine workflow involving Climate Quiz and an ontology learning algorithm

Task Structure. As depicted in Figure 2, Climate Quiz invites Facebook users and their online friends to evaluate whether two concepts presented by the system are related (e.g. "environmental activism", "activism"), and which label is the most appropriate to describe this relation (e.g. "is a sub-category of"). Similarly to Verbosity and Common Consensus, the system controls the types of relations between concept pairs, but our consideration set contains both generic ("is a sub-category of", "is identical to", "is the opposite of") and domain-specific ("opposes", "supports", "threatens", "influences", "works on/with") relations. Two further relations, "other" and "is not related to" were added for cases not covered by the previous eight relations. The game's interface allows players to switch the position of the two concepts or to skip ambiguous pairs.

Incentive Scheme and Dissemination. Similarly to RISQ! and PhraseDetectives (Poesio et al, 2013), Climate Quiz leverages the potential of social networking systems, particularly Facebook, for attracting players. Built-in notification systems (top right corner of the interface) and real-time progress statistics ("Level status" section) help engage Facebook users.

Participants earn one point for each matching answer, but can also lose points if their opinion differs from the majority of players. If in doubt, the system awards a point in order not to discourage players – if the first user selects relation A, for example, and the second user selects B, both receive a point since a majority solution has yet to be determined. If the first two players have answered A, however, the answer of a third player who does not agree with them will be considered wrong. This inevitably biases the baseline towards an early majority. This approach does not affect the gathered collective intelligence since i) the baseline results are only revealed to the player once he has answered the question and ii) the final result is computed on all answers independently of the order in which they were given.



Game Notifications

Correct. 100% of players agreed

Points: +1

Top Scores April 2012

1	Ernest Rudel Level: 8	100
2	Matt Dearborn Level: 6	61
3	Tigre Timide Level: 5	42
4	Thomas Brudertom Jr. Level: 4	21
5	Arno Scharl Level: 4	18
6	Jan Peters-Anders Level: 1	2
7	Florian Winter Level: 1	2

The „Climate Quiz" is a Facebook application in the tradition of Games with a Purpose. By playing the game, you contribute to a scientific project that aims to raise awareness and share environmental knowledge. You will face two types of challenges: (i) select correct relations between two concepts; (ii) answer climate-change related questions. <more>

Friends Invited: 3 April 2012 Bonus Points: 0

You will earn **10% of the points a friend makes** after accepting your invitation! The calculation is recursive, so if this friend invites another one you will even get more bonus points.

Figure 2: The Climate Quiz user interface

A possible disadvantage is that the current approach might discourage players in cases where they lose points for a correct answer. To mitigate this problem, we pre-tested a large portion of the game input data with scientists and other climate change experts. Collecting answers from trusted players before the public launch of the application increases the quality of the baseline and thereby improves game experience and player motivation. In the future, we will experiment with grading schemes that award points retrospectively only when a decision about a task has been made, similarly to the approach taken in the PhraseDetectives game.

Participants are given immediate feedback about each answer in terms of the percentage of players who agreed/disagreed with their decision as well as the majority voted relation if the player's answer differs from it (top right corner of the interface). This feedback constitutes a continuous training mechanism through the game and increases transparency by explaining how the points are provided.

To attract a sufficient number of players, we used a combination of press releases, presentations (e.g. at an online conference organized by the World Bank's Connect4Climate initiative), paid Facebook ads, and personal networking. To maintain and grow the resulting community of players, incentives include a levelling system with the opportunity to unlock additional game features, the comparison of a player's performance vis-à-vis the network of online friends, and a leader board showing monthly scores and progress statistics.

Implementation. Climate Quiz builds upon and extends the social application framework of Sentiment Quiz (Rafelsberger & Scharl, 2009), a publicly available Facebook application released as part of the US Election 2008 Web Monitor (Scharl & Weichselbraun, 2008) and collecting the political opinion of Web users as well as sentiment lexicons to support sentiment detection algorithms. Climate Quiz was implemented in PHP and JavaScript using the jQuery Framework and Facebook's Graph API. Using the Graph API enables a deep integration with the Facebook user experience and allows retrieving information about the user for further statistical analysis. Climate Quiz uses a Model-View-Controller (MVC) design pattern to allow easy maintenance and extensibility.

4. CLIMATE QUIZ EVALUATION

Climate Quiz was launched on 18 April 2012, together with a dedicated Facebook community page to assist the dissemination process. The game's input data consisted of environmental concept pairs extracted by the ontology learning algorithm from Anglo-American news media coverage between January and December 2011.

In this section we provide an evaluation of the game results obtained until October 2012. We focus on evaluating two aspects of the game. Firstly, we assess the success of the game in terms of its usage statistics (Section 4.1). Secondly, we analyse the quality of the obtained results (Section 4.2). For both evaluations we aimed to compare our work to previous efforts, but were hampered by two main obstacles. Firstly, there is a lack of widely accepted (and approved) evaluation measures for games, although a few measures have emerged over the years as *de facto* such as the game's *throughput* and the *average lifetime play* introduced in von Ahn's (2008) seminal work. Many papers describing games, however, still do not report on these measures. Secondly, existing games approach diverse tasks, have different setups, different goals and time-spans (some have been running for years, others only reached a prototype level and were evaluated for short test runs). Therefore, the provided comparisons are only informative but still useful to discover strengths and weaknesses of Climate Quiz.

4.1. Evaluation of Game Usage

In this section we measure aspects related to the game's usage and compare our results to findings published in the literature (summed up in Table 3), including usage statistics of the knowledge acquisition approaches described in Section 2.2. as well as those of two games for language resource acquisition detailed in (Chamberlain et al, 2013).

Number of players. Within the first week, a total number of 275 users had played the game, generating 7,836 ontology relations and 1,563 quiz answers. Of these 275 players, 222 had become returning visitors; 171 (120) of them returned for at least five (ten) times. Until the end of October, the total number of players had increased to 648, yielding 19,896 ontology relations and 3,871 quiz answers. Of these 648 players, 532 have become returning visitors; 409 (310) players returning at least five (ten) times. Among the knowledge acquisition games, Climate Quiz has been tested on the largest player population, which, however still lags behind the 2000 and 2700 players attracted by PhraseDetectives (PD) and JeuxDeMots (JDM) respectively (Chamberlain et al, 2013). Note, however, that PD and JDM have been running for much longer periods of time (32 and 56 months respectively), so Climate Quiz could eventually reach similar levels of players with time.

	Total Players	Players/ Month	ALP (min)	MPT (min)	Gender Ratio	Avg. contrib. per person
ClimateQuiz	648	92	10	60	48%F	25
Games for factual KA						
Verbosity	267	--	24	180	--	29 facts
Categorilla	--	--	3	--	--	--
Free Association	--	--	2	--	--	--
WhoKnows?	165	--	--	--	--	--
RISQ!	118	--	--	--	--	--
OntoPronto	270	--	--	--	11%F	--
UrbanMatch	54	--	3.17	--	--	--
BetterRelations	359	--	7	--	--	41
SpotTheLink	16	--	--	--	--	23.5
Games for linguistic KA						
PhraseDetectives	2000	62	35	--	65%F	--
JeuxDeMots	2700	48	25	--	60%F	--

Table 3: An overview of ClimateQuiz usage statistics and their comparison with results from other games for factual and linguistic knowledge acquisition

Chamberlain and colleagues (2013) measure the success of advertising and the motivation to join the game by computing the average number of players recruited per month. In the 7 months in which the statistics above have been collected, 648 players have been recruited thus leading to a recruitment level of about 92 players per month. This is superior to the 62 player/month and 48 players per month reported for PhraseDetectives and JeuxDeMots, potentially thanks to the viral advertising mechanisms enabled by Facebook. No knowledge acquisition game reports on this statistic, as in fact most of these games have been running for less than a month.

Length of play. A good level of user engagement for Climate Quiz is reflected by a typical game session lasting an average of 10.33 minutes (average lifetime play – ALP). This value is superior to all values reported by other knowledge acquisition games, with the exception of Verbosity with an ALP of 24 minutes. In comparison, PD and JDM report significantly higher ALPs of 35 minutes and 25 minutes respectively, potentially because they both involve reading text snippets thus naturally requiring longer for performing a task. The maximum play time (MPT) was somewhat over an hour, so a third of the value achieved by Verbosity.

Gender Distribution. In terms of the demographic structure of the player base, the gender distribution shows 325 male and 303 female participants (20 participants did not provide the information in their Facebook profile). The statistics of PD and JDM suggest a high number of female players, leading to female ratios of 65% and 60% respectively. In contrast, Climate Quiz has attracted fewer females (48%). This is however superior to the 11% female ratio reported by OntoPronto, the only knowledge acquisition game that provides this measure.

Contribution distribution over players. On average, each Climate Quiz player contributed about 25 judgments in the game, which is roughly similar to the numbers reported by other knowledge acquisition games. Note, however, that previous crowdsourcing systems have shown that the distribution of contributions over their contributor base is typically uneven (or even biased), with a small number of high scoring contributors providing most data. For example, in PD the 10 top scoring players had 60% of the total points and made 73% of all annotations. In the Facebook version of the same game, the top 10 players had 89% of the total points and provided 89% of the total annotations. This uneven distribution is only somewhat verified for Climate Quiz: while the top 10 scoring players have earned more than half of the total points awarded by the system (54%), they contributed only 37% percent of the data, so a much lower percentage than in PD (Figure 3). Therefore, our dataset is less likely to be biased as it has a higher active contributor base, including around top 50 players who accumulated 88% of the total points and provided 70% of the data. The rest of the players provide 30% of the data.

Social Network. A side effect of deploying Climate Quiz over Facebook is the possibility of collecting "friend" links between players, an advantage over games deployed on stand-alone websites. This information can provide interesting insights into the social fabric of the player population. For example, Figure 4 uses a force directed layout based visualization to depict the social network of the players based on declared friend relations. It reveals a strongly connected core of players but also several "casual" players connected to only a few friends in the game. There exists also a set of players that are not connected to any other player, but these have been filtered out to maintain the clarity of the figure. This seems to be an ideal situation, in which a strong, motivated, self-reinforcing community has been built. This community is likely to further advertise the game and to provide contributions both because its members are interested in the domain, but also because of social motivation (e.g., performing better than friends). At the same time, casual players help reduce bias by bringing in diverse insights, assuming that they share fewer commonalities with the core players (different interests, location, and economic status). An in depth analysis of the social structures behind the game are out of the scope of this paper, but future work will correlate social network information with other aspects such as the quantity and quality of the contributions or the geographic location of the participants.

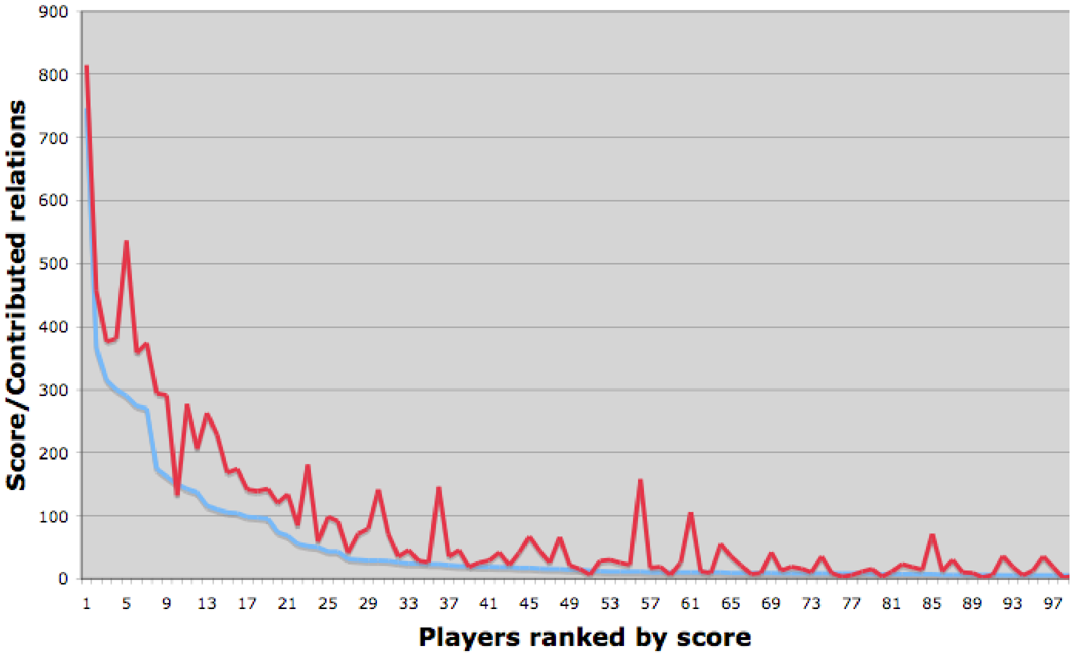


Figure 3: Distribution of player contributions (red/dark line) in comparison to their score (blue/light line)

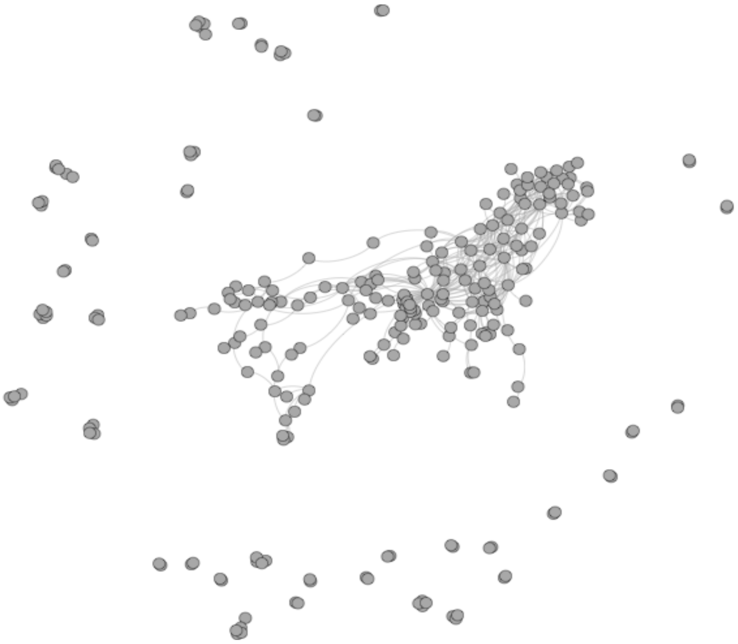


Figure 4: "Friend" relations between players (excludes players with no connections)

4.2. Evaluation of Task Throughput and Data Quality

A total number of 1,213 distinct concept pairs were assessed between April and October 2012. A final relation could be identified for 424 of these concept pairs. The other 789 pairs remain in the game and will require additional user interactions to confirm the majority opinion. In this section we assess the speed at which data is produced and estimate the quality of the data.

Throughput. The average speed at which players are answering is 19 seconds per ontology relation and 27 seconds per quiz question. We therefore estimate the throughput of the game, in terms of provided ontology relations per human hour, to about 180 relations per hour. This value is low when compared to those reported by other games, ranging from 350 to 648 contributions per hours (Table 4). This is an indication of the difficulty of the task at hand as we discuss next.

	Throughput	**Accuracy**
Climate Quiz	180 relations/H	72% (up to 95% for some relations)
Games for factual KA		
Verbosity	--	85%
Categorilla	396 guesses/H	--
Free Association	594 guesses/H	--
UrbanMatch	485 links/H	99.06%
BetterRelations	350 decisions/H	--
Games for linguistic KA		
PhraseDetectives	450Annot/H	84%
JeuxDeMots	648Annot/H	--

Table 4: Throughput and Accuracy of Climate Quiz and other GWAPS for factual and linguistic knowledge acquisition

Data Quality. Comparison against a gold standard dataset is an ideal approach to check the quality of the produced data set. However, in our case, there is no readily available gold standard for the problem of acquiring term relations in the environmental domain. Therefore, we compared the results obtained from the game with those generated by two colleagues, which were experts in knowledge representation, but not in the area of climate change. For all pairs involving climate change knowledge, the annotators have gathered the necessary information from online sources in order to provide a correct relation.

Table 5 provides an overview of the agreement between: 1) the two annotators; 2) the game and each individual annotator; 3) the game and a gold standard dataset derived from the relations agreed upon by both annotators. We measure an overall agreement for all these cases as the ratio between all agreed relations by the two compared parties and all relations in the dataset (i.e., 424). To shed light on the difficulty level of the individual relations, we compute relation specific agreements. For a relation type R and two annotators A and B, the specific agreement is $2*R_{A\&B}/(R_A+R_B)$, where $R_{A\&B}$ is the number of pairs for which both A and B agree that a relation R holds, while R_A and R_B are the number of pairs judged as related by R by annotator A and B respectively.

A first conclusion is the high difficulty level of the task at hand: our two annotators, despite their knowledge representation background and careful consideration of domain specific term pairs, only agreed on the relations assigned to 35% of the pairs. In these conditions, although the pair wise agreements between the game and the individual annotators are rather small, they are actually superior to the annotator agreement, thus showing that the quality of the game results is similar to what one could hope for from another annotator. We created a gold standard dataset from the relations on which both annotators agreed thus excluding highly ambiguous cases. The agreement of the game data with the gold standard was considerably higher, 72%, reaching relation specific agreements of even 95%. These values are in line with accuracy values reported by other games and summarized in Table 4.

	Ann1<>Ann2	**CQ <> Ann1**	**CQ <> Ann2**	**CQ <> GS**
Is a sub-category of	66%	61%	61%	84%
Is identical to	50%	12%	19%	40%
Is not related to	29%	40%	41%	71%
Is the opposite of	0	0	0	0
Other	35%	4%	4%	11%
Influences	11%	0	40%	0
Opposes	23%	32%	44%	55%
Supports	16%	24%	35%	59%
Threatens	60%	71%	77%	95%
Works on/with	10%	16%	49%	25%
Overall Agreement	**35%**	**38%**	**47%**	**72%**

Table 5: Relation specific agreements and overall agreements between annotators and game results (CQ: Climate Quiz, Ann: Annotator, GS: Gold Standard)

Taking a more detailed look at the evaluation data, we discovered the following causes of the high difficulty level of the task:

- **Some relations are more controversial than others**. The relation specific agreements in Table 5 suggest that some relations are easier to assign than others. For example, *subsumption* reached high levels of consensus between all parties, including 84% agreement rate with the gold standard. The *threatens* relation seems to be the least controversial among the domain specific relations reaching agreement levels of over 60% and even 95% in the gold standard comparison. At the other end of the spectrum are relations for which there is very little (or even no agreement), namely *is the opposite of* and *influences*.
- **Too many relations to choose from**. Players are asked to assign one of the 10 relations that potentially cover most relations between the input terms. However, this large number of relations can potentially confuse players and increase the cognitive load of the game. Indeed, experience from expert-based annotation (Hovy10a) has shown that annotators should not be asked to choose from more than 10, ideally 7, categories. Compared to these guidelines, crowdsourcing tasks typically present much fewer choices for classification style problems, in most cases ranging between 2 (binary choice) and 4 categories. Some authors justify this reduction of categories as a means to make the task simple enough to be amenable to be solved by non-expert annotators. For example, Snow et al. (2008) reduce an event ordering task from 14 to two relations. Experimental results on MTurk support this finding: as the number of choices increases, annotation quality deteriorates (Fort et al., 2011; Hong and Baker, 2011). Future versions of the game should reduce the number of offered relations, especially those that proved controversial.
- **Overlapping relation semantics**. We also observed that some relations have somewhat overlapping semantics, with multiple relations fitting the same pair of terms. For example, some pairs termed with the relation *influences* could also be termed with *supports* or *threatens* - *(climate change, glaciers), (global warming, environment).*
- **Ambiguous and obscure terms**. Since the input terms are generated automatically by the ontology learning algorithm, in some cases they are ambiguous or obscure making the assignment of the relations very difficult. Examples are terms such as “more acidic” or “climate biz”.

4.3. Evaluation Summary

Based on the evaluations above, we conclude that while Climate Quiz has attracted a significant number of players (the highest number of all knowledge acquisition games) and managed to build a 50+ core community of players (as opposed to only 10 in PD), it has achieved only medium average lifetime play values. Additionally, its throughput was the lowest of all games and so was the agreement of the game results with the gold standard dataset.

The evaluation also revealed the high difficulty of the task (Section 4.2) which we assume to be the main cause of players playing the game for short intervals only (hence the average ALP) and providing results that have a low quality when compared to other games (although, in line with the quality provided by paid annotators). More specifically, we distinguish two core problematic issues that lead to the limitations of the game.

Firstly, the game is fed *noisy input data*, generated automatically by the ontology learning algorithm and containing terms that are ambiguous, obscure or do not make sense at all. A severe negative effect is that such confusing terms frustrate players and reduce the enjoyment of the game, which is the main motivational factor Climate Quiz relies on. Therefore, frustrated players play less (lower ALP) and are likely to lose motivation and leave the game, thus preventing the game for maintaining a stable community over long periods of time and jeopardizing its long-term success. Noisy input data also leads to wasting precious game resources (i.e., players' time and effort) on obscure terms and inherently to low performance as disagreement tends to be high on these ambiguous pairs.

Secondly, the game *loses good quality output data*. As discussed in Section 4.2 above, the high number of relations to choose from and their semantic overlap often lead to cases when a pair of terms can be correctly related with multiple relations (e.g., threatens and influences), Climate Quiz, however, derives a single relation between any input pair using a majority voting based mechanism causing that less popular but still correct relations are excluded from the final result set. For example, the game assigned the relation *works on/with* to the pair (*green industry, clean energy products*) as the most popular one, and therefore did not include the relation *supports*, which was the second most popular relation voted by the players and can be considered a correct relation.

5. HYBRID-GENRE CROWDSOURCING WORKFLOWS

As a way to mitigate the problematic issues discussed above, we propose a workflow that combines two different crowdsourcing genres in order to leverage their complementary strengths, hence the term *hybrid-genre workflow*. In our context, and considering the pros and cons of crowdsourcing genres discussed in Section 2.4, we assign simple (and boring) tasks to micro-workers and keep more complex (but interesting) tasks for game players thus ensuring game enjoyment and reinforcing players' intrinsic motivation. Therefore, our workflow is novel compared to the workflow types described in Section 2.3, which either relied on a single crowdsourcing genre (most frequently mechanised labour) or combined machine and human computation. Concretely, our workflow has three stages (see Figure 5).

- **Stage 1: Judge Pair Relatedness**. This stage addresses the problem of noisy input data by asking CrowdFlower workers to check which pairs of terms extracted by the ontology learning algorithm might be related before feeding these into the game. Acting similarly as the "Find" phase of the Soylent workflow (Bernstein et al, 2010), this stage detects the problem instances worth investigating and therefore reduces the ambiguity of the input data. We hypothesize that this will lead to several positive effects such as (i) a more enjoyable game resulting in higher player motivation and retention as well as (ii) higher quality game results in terms of better agreement with the gold standard.

- **Stage 2: Assign Relation**. Climate Quiz is used in this stage to solve the complex problem of assigning one of ten relations between term pairs resulting from Stage 1. As such it corresponds to the "Fix" phase of the Soylent workflow which solves the problem instances identified in the previous Find phase.

- **Stage 3: Check Relation Correctness.** This stage asks workers to assess the correctness of the relations assigned in stage 2 above (similarly to Soylent's "Verify" stage). As such, it should further increase the quality of the game's output but also extend it with potentially correct but rejected relations thus alleviating the problem of losing good quality output data.

We have verified the feasibility and the positive effects of such a workflow by implementing the mechanised labour stages (2 & 3) and testing them with the Climate Quiz data obtained between April-October 2012. We describe the implementation of these stages as well as their evaluation in the next subsections.

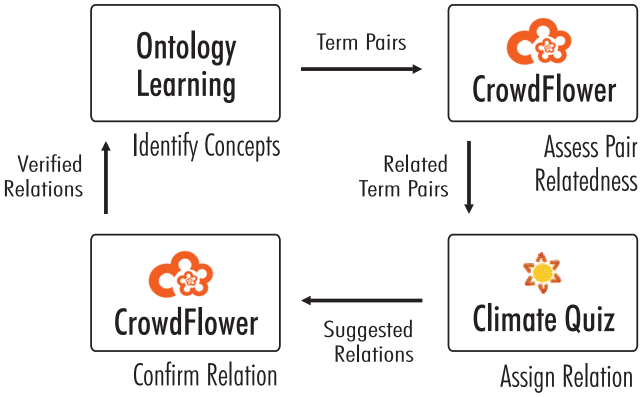


Figure 5: Hybrid-genre workflow

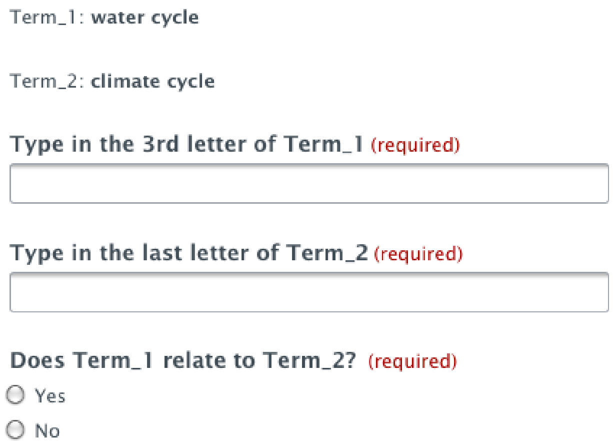
5.1. Stage 1: Judge Pair Relatedness

The first stage aims to clean the input data for the game by requiring crowdworkers to evaluate whether a relation might exist between the two terms. The importance of clean input data to maintain player motivation and ensure high quality results has been identified by other games as well (e.g., PhraseDetectives, GuessWhat?!, RISQ!), however, the cleaning process has been implemented primarily manually and outside of the crowdsourcing cycle.

Input Data. The input data consisted in the 424 term pairs from the relation set resulting from Climate Quiz.

Task Design. We used the CrowdFlower mechanised labour platform for our experiments. The interface of the task was created with the platform's Builder tool. Each task (or unit in CF terminology) presented the workers with two terms and asked them to vote whether a relation between the two terms exists (see Figure 6).

While various techniques exist to filter out invalid responses after the completion of a crowdsourcing task, it is preferable to reduce malicious behaviour as much as possible on the first place. Task interface design plays a key role here: indeed, both (Laws et al, 2011) and (Kittur et al, 2008) have extended their interfaces with explicitly verifiable questions which forced workers to process the content of the task and also signal to them that their answers are being scrutinized (e.g., asking the workers to type in a word from the processed document or the number of references that a Wikipedia article has). This seemingly simple technique has increased classification accuracy for (Laws et al, 2011) to 75% and reduced the percentage of invalid responses to only 2.5% for (Kittur et al, 2008). Therefore, we have also included two verification questions in the tasks that could only be answered correctly if the workers actually read the two terms. One question requested workers to type in the 3rd letter of the first term, while the other required them to provide the last letter of the second word. Because their primary role is to get workers to look at the task content on the first place, verification questions are usually simple (e.g., in the case of Kittur counting the number of references, images and sections in the input Wikipedia article). In our case, the task data is small (two terms) thus further restricting the complexity of explicitly verifiable questions that can be asked in relation to these two terms.



Term_1: water cycle

Term_2: climate cycle

Type in the 3rd letter of Term_1 (required)

Type in the last letter of Term_2 (required)

Does Term_1 relate to Term_2? (required)

○ Yes

○ No

Figure 6: User interface for the pair relatedness task

We relied on training micro-workers as an additional method to ensure the quality of the results. Firstly, we provided detailed instructions of how to perform the task including many examples of related and non-related terms. Secondly, we augmented the input data (424 pairs) with the recommended 5% of gold units, that is 22 units. We provided 11 positive and 11 negative examples. CrowdFlower uses these gold units to train the workers on the go but also to detect low-performing players and to exclude their work automatically from the final result.

Three workers judged each unit, and the total cost of the entire job was $4.66. The job (i.e., the collection of all our units) was completed within two hours. To ensure that the workers had the command of English necessary for completing the task, our job was made available only to workers from the UK and the USA.

Evaluation. Since it is based on the consensus of the two annotators, the gold standard dataset provides a good baseline for evaluating the precision with which micro-workers were able to predict whether a relation between a pair of term exists. From the 147 relations of the gold standard, 96 were judged as existing relations. However, 3 of these relations were of type "is not related to" and 9 were of type "other" and therefore should have been rated as non-existent. The remaining 51 relations were judged as non-existent. However, 24 or these relations are "is not related to" and 8 are "other" in the gold standard, so these were correctly judged. Therefore, the precision of the crowdsourcing method on predicting relatedness is ((96 - 12) + 32)/147 => 79%.

To evaluate the effectiveness of this method on more ambiguous pairs, we also computed its precision on sets of pairs with decreasing level of ambiguity. For example, on the 105 relations on which both annotators as well as the game results agree, the precision of the method was 83%. For more ambiguous sets of relations this value decreased to 59% for the pairs on which only two parties agree and 61% on the set of pairs on which no parties agree. In all our calculations we considered a "non-existent" rating by the micro-workers correct is at least one of the two annotators or the game rated the corresponding relation as "is not related to", and incorrect otherwise. Conversely, we considered an "existent" rating by the micro-workers incorrect if at least one of the two annotators or the players rated the corresponding relation as "is not related to", and correct otherwise. We can therefore conclude that the method has a good precision in predicting the existence of relations in non-ambiguous cases (79% - 83%), but works also well with ambiguous cases reaching precisions of about 60%. We believe that this reduction of ambiguous terms will lead to a higher enjoyment of the game. This hypothesis can only be evaluated in the future when the mechanised labour and game stages will be tightly coupled.

Our second hypothesis was that a less ambiguous input data set could lead to improved game result quality. To check this hypothesis, we removed all pairs judged as non-related by the micro-working stage from the Climate Quiz output and recomputed its agreement with the gold standard. We obtained an agreement of 76%, thus resulting in an improvement of 4% over the 72% baseline agreement level, when the input data was not cleaned.

5.2. Stage 3: Check Relation Correctness

Although the exclusion of the ambiguous pairs has already led to improvements in the quality of the output data, we added an additional crowdsourcing step after the game in which we asked micro-workers to vote whether a relation is correct or not. Our hypothesis was that this would allow improving the results in two ways:

- **Improved precision** could be obtained by workers verifying the output of the game and judging which relation is correct or not.
- **Extended result set**: We use crowdsourcing to evaluate some of the “rejected” relations in order to find those that could also be included in the final result set and therefore broaden the scope of the ontology built by the ontology learning algorithm.

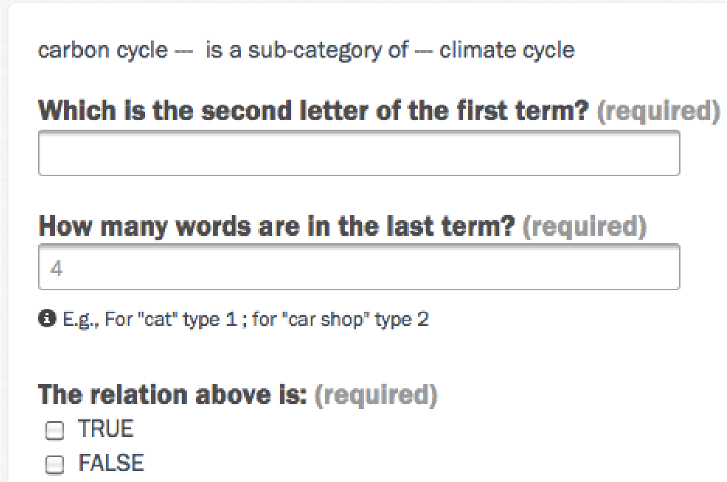
Input Data. From the total set of 424 relations produced by Climate Quiz we excluded those judged as unrelated by our previous crowdsourcing task as these would have not made it to the game if such a workflow would be in place. We then extended the resulting set of 255 relations with 45 rejected relations (i.e., that did not make it to the game’s output). In order to select the most promising ones we computed a metric to estimate the support for a given relation R ($support_R$) as directly proportional to the votes for this relation ($vote_R$) and inversely proportional to the sum of votes for all the relations that were identified between that pair (including the cases when the pair was skipped: $allVotes_{pair} + skip_{pair}$) and the number of relations identified between a pair ($nrRel_{pair}$):

$$support_R = vote_R/((allVotes_{pair} + skip_{pair})*nrRel_{pair})$$

To assess the effectiveness of this metric in predicting the support for a relation, we computed its average and median values for three partitions of the dataset: (i) the relations on which no agreement was reached (avg = 0.11, median = 0.045); (ii) the relations on which two parties agreed (avg = 0.17, median = 0.077); and (iii) the relations on which all parties agreed (avg = 0.33, median = 0.24). We observe that, the higher the support for a relation the more likely that it is correct. Therefore, we selected the rejected relations that had the highest support values. This yielded a total of 300 relations. We provided 15 gold units.

Task Design. Each task displays a triple (the two terms and their relation) and asks workers to judge whether the relation is correct (TRUE) or not (FALSE). As with our previous experiment, we have also added two verification questions one asking for the second letter of the first term and the other requesting the number of words in the second term (see Figure 7). The interface also contained detailed examples of how to judge the correctness of a relation, including examples of both correct and incorrect relations.

Crowdsourcing settings. Given the higher difficulty of this task, we increased the number of judgments/unit to 5 from the previous 3. We offered the job to workers from the US alone. We grouped 5 units per page and paid $0.02 per page, leading to a total cost of $10.98. The job finished in about 11 hours.



carbon cycle — is a sub-category of — climate cycle

Which is the second letter of the first term? (required)

How many words are in the last term? (required)

4

E.g., For "cat" type 1 ; for "car shop" type 2

The relation above is: (required)

□ TRUE

□ FALSE

Figure 7: User interface for the relation verification task

Evaluation. The evaluation focused on verifying the two hypotheses related to the usefulness of this crowdsourcing stage.

Improved precision. We removed all the relations considered FALSE by the micro-working stage from the game's output (i.e., from the 255 relations that were considered existent by the first crowdsourcing stage) and recomputed the agreement with the gold standard data. This lead to an agreement of 78%, thus another 2% increase from the agreement levels obtained when introducing the first crowdsourcing stage (76%) and a total of 6% increase from the scenario where no crowdsourcing was used. Therefore our hypothesis of improving the precision of the overall output of the process is verified.

To get an understanding of the precision of the crowd-workers, we compared their ratings against the 72 relations on which all parties agreed. We found that the crowd considered only 2 relations as FALSE, namely: *(Intergovernmental Panel on Climate Change, works on/with, assessment report)* and *(sea level rise, is not related to, Himalayan glaciers)*. However, in both cases the confidence in the joint judgment of the micro-workers was not absolute, but had low levels of 0.62 and 0.8 respectively. Therefore, in the case of this non-ambiguous set of relations, the precision of the method is very high, namely 97%.

We also observed that the confidence values assigned by CrowdFlower to each judgment correlate well with the ambiguity of the relations. Indeed, for the pair set on which none of the parties agreed, the average CF confidence was 0.83; on the pairs on which two parties agreed it was 0.84; while for the pairs on which all agreed the average confidence was 0.93. In future work we will investigate methods that use these confidence values, for example, by accepting only those judgments that have a confidence level over a certain threshold.

Extended result set. From the 45 rejected relations 22 were judged as FALSE and 23 as TRUE. Since we had no evaluation baseline here, one of our annotators manually evaluated whether the assignments by the crowd were correct. We found that only 2 out of the 22 FALSE judgments were incorrect, that is, they were actually correct relations. 15 of the TRUE judgments were incorrect. Therefore, the precision on this dataset was of 63%.

We conclude therefore, that in its current form this stage would introduce considerable noise into the final results when judging rejected relations. We see two ways of making use of it. Firstly, it could be used as a way to filter the "rejected" output and present the results judged as correct to a human annotator for further verification. Secondly, a more sophisticated selection mechanism could be used which would favour only high confidence judgments. We suspect this is possible because we observed a high correlation between confidence values and the correctness of the relations according to the annotator: those relations which were judged TRUE correctly, had a much higher average confidence (0.89) than those which were judged TRUE incorrectly (0.76). This remains however future work.

One interesting observation from this experiment was that, with sufficient examples provided through instructions and gold units, micro-workers were capable of understanding even rather subtle knowledge representation concepts, such as the meaning of subsumption between a more specific concept (the first term) and a more generic concept (the second term). For example, they correctly identified 8 incorrect cases where the more generic concept was first, for example: (*climate oscillation, is a sub-category of, Antarctic oscillations*) or (*deniers, is a sub-category, climate deniers*). This distinction was difficult for game players, and they often assigned subsumption wrongly without taking into account the order of the terms.

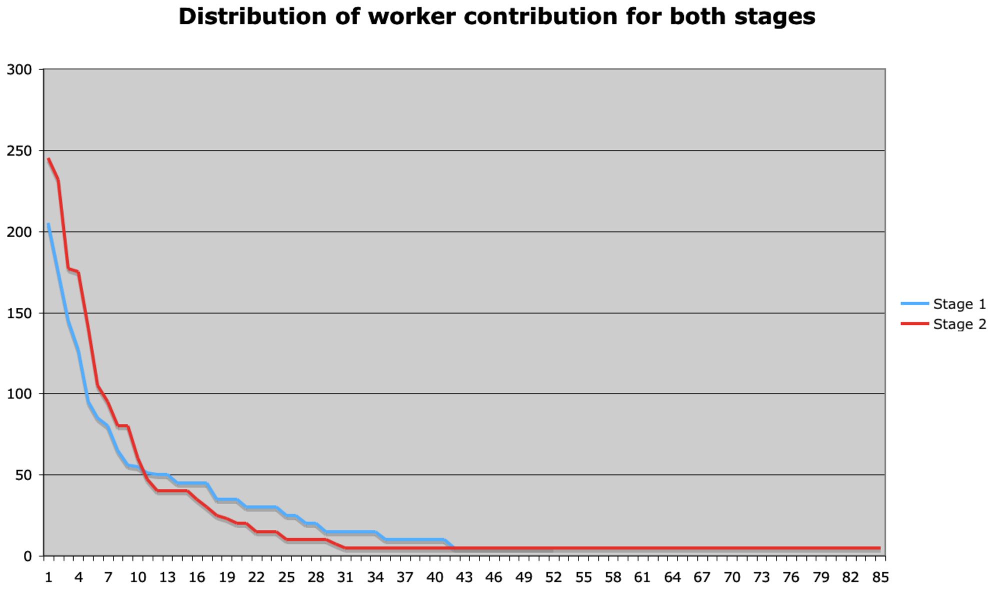


Figure 8: Distribution of worker contributions in both mechanised labour stages

5.3. Overall Evaluation and Genre Comparison

We extend the evaluations of the mechanised labour stages from Section 5.1 and 5.2 with additional measures similar to those discussed for Climate Quiz in Section 4. These are summarised in Table 6 and compared against the Climate Quiz evaluation.

Firstly, the evaluation of usage statistics similar to the one performed in Section 4.1., revealed that, in terms of number of participants 53 workers contributed to stage 1 and 85 to stage 2. Other participant related data such as gender ratio or the social network between workers are not available from CrowdFlower. In terms of the contribution distribution over the workers, a similar pattern emerges as in the case of Climate Quiz, where a few top-workers provide most of the judgments (see Figure 8). However, while in the case of the game the top 10 players provide 37% of all judgments, this percentage is much higher in the case of mechanised labour and amounts to 56% for stage 1 and 65% for stage 2. Therefore, in this case, the mechanised labour stages have more tendency towards bias, partially also due to their shorter time frame. The throughput of the mechanised labour methods amounts to 734 judgments/H for stage 1 and 166 for stage 2. This difference in throughput demonstrates that more complex tasks, e.g., judging whether a relation is correct as opposed to just estimating whether it exists, take more time.

	CrowdFlower Stages	**Climate Quiz**
Number of participants	53 (stage 1), 85 (stage 2)	648
Cost		
Price per task	$0.15	$0
Speed		
Set-up time	One week	2 months
Time for experiments	2 H (stage 1), 11 H (stage 2)	6 months
Throughput (judgments/H)	734 (stage1), 166 (stage 2)	180
Throughput predictability	experiments finished within hours	completion difficult to estimate
Data Quality		
Maintaining motivation	no effort to recruit micro-workers	Significant effort for recruiting and maintaining players
Incentive to cheat	30% of the judgments discarded as unreliable probably due to cheating	NA
Task complexity	simple, binary classification tasks	complex task of choosing between 10 relations
Importance of task interestingness	Micro-workers solve all tasks	players skip tasks and leave game due to ambiguous terms
Contributions of top 10 workers/players	56% (stage 1), 65% (stage 2)	37%

Table 6: Main evaluation statistics for CrowdFlower in comparison to Climate Quiz

By contrasting the mechanised labour statistics against those of the game, we notice that they support most genres differences predicted in Table 2. In particular, as shown in Table 6, the null unit prices for the game rival the average $0.15 unit price on mechanised labour platforms. This advantage of the games is however offset by longer (and costlier) set-up times, slower execution times and throughput as well as the difficulty to estimate completion times due to the uneven availability of players over time. In terms of data quality, it is difficult to directly compare the precision of game and mechanised labour approaches as they focused on the execution of different tasks. We can however draw a few conclusions about issues related to data quality including: (i) maintaining player motivation requires significant effort for games while it is not an issue on crowdsourcing platforms; (ii) workers performed simple tasks while players managed much more complex tasks; (iii) players often skipped those tasks that involved ambiguous terms; (iv) mechanised labour results tend to have a higher bias than those obtainable through games. In terms of the incentive to cheat we cannot provide a conclusion as this phenomenon was not investigated within Climate Quiz, but it is known that CrowdFlower rejected 30% of the contributions potentially obtained as a side effect of cheating.

6. SUMMARY AND FUTURE WORK

In this paper we investigated several aspects related to the use of crowd-based social collaboration platforms for knowledge acquisition. Our survey of approaches for acquiring knowledge assets relevant for the Semantic Web has shown that the game-based genre is the most popular, although a trend of moving towards mechanised labour platforms is currently taking place. The approaches we overviewed made no or very limited use of workflow mechanisms, although several different types of workflows have been extensively studied in other fields such as NLP. None of the workflow mechanisms we know about, neither in the knowledge acquisition nor in other fields, combine different crowdsourcing genres although there is abundant evidence from literature on the high complementarity of these genres.

We described Climate Quiz, as a typical example of a game for knowledge acquisition in the environmental domain. Our evaluation of this game and its comparison to other similar games has shown that Climate Quiz had the highest amount of players among knowledge acquisition games and that it has successfully established a healthy core community of players (50+). The evaluation also revealed average player involvement (low ALP), low throughput and sub-optimal data quality. These were linked to the games' noisy input data and complex task structure.

Starting from the need to solve the limitations of Climate Quiz, we proposed using a workflow that, for the first time according to our survey in Section 2.3, combines two crowdsourcing genres to solve one problem. The workflow leverages the extrinsic motivation of the crowd-workers, and their ability to perform simple (and boring) tasks in order to filter the input data and verify the output data of the game. This allows the game to present players with mostly interesting problem cases thus providing a more engaging gaming experience and maintaining their intrinsic motivation.

We envision several lines of future work. Firstly, we will explore in more depth the social relations of the Climate Quiz players and investigate whether they can be used for building advanced filtering and player profiling mechanisms (e.g., assign trust levels per groups of interconnected players). Secondly, we will couple the developed mechanised labour tasks with the Climate Quiz game to benefit of the expected improvements. This will also involve developing more sophisticated mechanisms of interpreting the CrowdFlower output, in particular, by making better use of the provided confidence scores. Thirdly, we will experiment with other types of hybrid-workflows. The current workflow has primarily benefited from the complementarities of the two genres related to data quality. We envision other types of workflows that might focus on optimizing the speed or the cost of a knowledge creation project, e.g., micro-workers could supplement games in periods when their throughput is low, thus ensuring that the overall throughput of the workflow remains constant over time. Last but not least, as part of the uComp project (www.ucomp.eu), we will build a generic crowdsourcing infrastructure capable of building and managing such hybrid-genre workflows.

Acknowledgement

We thank Adrian M.P. Brasoveanu and Stefan Gindl for their help with creating the gold standard annotation and visualizing the social network connecting the players. Section 2 partially relies on earlier work performed together with Kalina Bontcheva. The Climate Quiz Facebook application was developed within DIVINE (www.weblyzard.com/divine) and Triple-C (www.modul.ac.at/nmt/triple-c), two research projects funded by FIT-IT Semantic Systems of the Austrian Research Promotion Agency and the Austrian Climate Research Program of the Austrian Climate and Energy Fund. The crowdsourcing methods were developed and evaluated as part of uComp (www.ucomp.eu), a European research effort in the CHIST-ERA program line.

REFERENCES

1. von Ahn, L., Mihir, K. & Blum, M. (2006). Verbosity: A Game for Collecting Common-Sense Facts. In R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, and G. Olson (Eds.) *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 75-78). ACM.

2. von Ahn, L. & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM, 51(8)*, 58–67.

3. Berners-Lee, T. Hendler, J. & Lassila, O. (2001) The Semantic Web. *Scientific American 284(5)*, 34-43.

4. Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D. & Panovich, K. (2010). Soylent: A Word Processor with a Crowd Inside. In *Proceedings of the 23rd ACM Symposium on User Interface Software and Technology,* (pp: 313 - 322).

5. Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Niraj Aswani, N. & Gorrell, G. (2013) GATE Teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation.* DOI: 10.1007/s10579-013-9215-6.

6. Brew, A., Greene, D. & Cunningham, P. (2010) Using Crowdsourcing and Active Learning to Track Sentiment in Online Media. In *Proceedings of the European Conference on Artificial Intelligence* (pp: 145–150).

7. Callison-Burch, C. (2009) Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp: 286–295).

8. Celino, I., Contessa, S., Corubolo, M., Dell'Aglio, D., Della Valle, E., Fumeo, S., & Krüger, T. (2012) Linking Smart Cities Datasets with Human Computation - The Case of UrbanMatch. In *Proceedings of the International Semantic Web Conference* (pp:34-49).

9. Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M. & Poesio, M. (2013) Using Games to Create Language Resources: Successes and Limitations of the Approach. In I. Gurevych & K. Jungi (Eds.) *The People's Web Meets NLP. Collaboratively Constructed Language Resources*. Springer.

10. Chklovski, T. (2005) Collecting Paraphrase Corpora from Volunteer Contributors. In *Proceedings of the 3rd International Conference on Knowledge Capture* (pp:115–120).

11. Dai, P., Mausam & Weld, D.S. (2010). Decision-Theoretic Control of Crowd-Sourced Workflows. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp: 130-135). AAAI.

12. Demartini, G., Difallah, D. E. & Cudré-Mauroux, P. (2012). ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-Scale Entity Linking. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 469-478). ACM.

13. Doan, A., Ramakrishnan, R. & Halevy, A. Y. (2011). Crowdsourcing systems on the World-Wide Web. *Communications of the ACM, 54(4)*, 86-96.

14. Eckert, K., Niepert, M., Niemann, C., Buckner, C., Allen, C., & Stuckenschmidt, H. (2010). Crowdsourcing the Assembly of Concept Hierarchies. In *Proceedings of the 10th Annual Joint Conference on Digital libraries* (pp.139-148). ACM.

15. Feigenbaum, E. A. (1977) The Art of Artificial Intelligence: Themes and Case Studies of Knowledge Engineering. In *Proceedings of the 5th International Joint Conference of Artificial Intelligence* (pp: 1014–1029).

16. Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics, 37(2),* 413-420.

17. Heath, T. & Bizer, C. (2011) Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool.

18. Hees, J., Roth-Berghofer, T., Biedert, R., Adrian, B. & Dengel, A. (2011). BetterRelations: Using a Game to Rate Linked Data Triples, In J. Bach and S. Edelkamp (Eds.) *Proceedings of the 34th Annual German Conference on Artificial Intelligence* (pp: 134-138). Springer.

19. Hoffmann, L. (2009) Crowd Control. *Communications of the ACM, 52(3)*, 16 –17.

20. Hong, J. & Baker, C. F. (2011) How Good is the Crowd at "real" WSD? In *Proceedings of the 5th Linguistic Annotation Workshop* (pp: 30–37).

21. Hovy, E. (2010) Annotation. In *Tutorial Abstracts of ACL*.

22. Howe, J. (2009) Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business. http://crowdsourcing.typepad.com/.

23. Kawrykow, A., Roumanis, G., Kam, A., Kwak, D., Leung, C., Wu, C., Zarour, E., & Phylo players. (2012) Phylo: A Citizen Science Approach for Improving Multiple Sequence Alignment. *PLoS ONE, 7(3):e31362*.

24. Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp: 453-456). ACM.

25. Kittur, A., Smus, B. & Kraut, R. (2011). CrowdForge: Crowdsourcing Complex Work. In *Proceedings of the Conference on Human Factors in Computing Systems* (pp: 1801-1806). ACM

26. Laws, F., Scheible, C. & Schutze, H. (2011) Active Learning with Amazon Mechanical Turk. In *Proceeding of the Conference on Empirical Methods in NLP* (pp: 1546–1556).

27. Lieberman, H., Smith, D.A. & Teeters, A. (2007). Common Consensus: a web-based game for collecting commonsense goals. In *Proceedings of the Workshop on Common Sense and Intelligent User Interfaces held in conjunction with the 2007 International Conference on Intelligent User Interfaces IUI*.

28. Little, G., Chilton, L.B., Goldman, M. & Miller, R.C. (2010). TurKit: Human Computation Algorithms on Mechanical Turk. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology* (pp: 57-66). ACM.

29. Liu, W., Weichselbraun, A., Scharl, A. & Chang, E. (2005). Semi-Automatic Ontology Extension Using Spreading Activation, In *Proceedings of the 5th International Conference on Knowledge Management* (pp: 145-153). Springer.

30. Markotschi, T. & Völker, J. (2010). Guess What?! Human Intelligence for Mining Linked Data, In *Proceedings of the Workshop on Knowledge Injection into and Extraction from Linked Data at the International Conference on Knowledge Engineering and Knowledge Management (EKAW-2010).*

31. Negri, M., Bentivogli, L., Mehdad, Y., Giampiccolo, D. & Marchetti, A. (2011) Divide and Conquer: Crowdsourcing the Creation of Cross-Lingual Textual Entailment Corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp: 670–679).

32. Poesio, M., Kruschwitz, U., Chamberlain, J., Robaldo, L. & L. Ducceschi, L. (2013). Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation. *ACM Transactions on Interactive Intelligent Systems*. In Press.

33. Quinn, A. J. and Bederson, B. B. (2011) Human Computation: A Survey and Taxonomy of a Growing Field. In *Proceedings of Human Factors in Computing Systems* (pp: 1403–1412). ACM.

34. Rafelsberger, W. & Scharl, A. (2009). Games with a Purpose for Social Networking Platforms. In C. Cattuto et al. (Eds.) *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia* (pp: 193-197). ACM

35. Sarasua, C., Simperl, E., & Noy, N., F. (2012). CrowdMap: Crowdsourcing Ontology Alignment with Microtasks. In P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, and J. Euzenat (Eds.) *Proceedings of the 11th International Conference on The Semantic Web* (pp: 525-541).

36. Siorpaes, K. & Hepp, M. (2008). Games with a Purpose for the Semantic Web, *IEEE Intelligent Systems, 23(3),* 50-60.

37. Scharl, A. & Weichselbraun, A. (2008). An Automated Approach to Investigating the Online Media Coverage of US Presidential Elections, *Journal of Information Technology & Politics, 5(1)*, 121-132.

38. Scharl, A., Sabou, M. & Föls, M. (2012). Climate Quiz – A Web Application for Eliciting and Validating Knowledge from Social Networks. In G. Bressan and R.M.

Silveira (Eds.) *Proceedings of the 18th Brazilian Symposium on Multimedia and the Web* (pp: 189-192). ACM.

39. Snow, R., O'Connor, B., Jurafsky, D. & Ng, A. Y. (2008). Cheap and Fast—but is it Good?: Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp: 254–263).
40. Strohmaier, M., Helic, D., Benz, D., Körner, C. & Kern, R. (2012). Evaluation of Folksonomy Induction Algorithms. *ACM Transactions on Intelligent Systems and Technology, 3(4)*, Article 74.
41. Thaler, S., Simperl, E. & Siorpaes, K. (2011) SpotTheLink: A Game for Ontology Alignment. In *Proceedings of the 6th Conference for Professional Knowledge Management* (pp: 246-253).
42. Thaler, S., Simperl, E. & Wölger, S. (2012). An Experiment in Comparing Human-Computation Techniques. *IEEE Internet Computing, 16(5),* 52-58.
43. Tudorache, T., Nyulas, C.I., Noy, N.F. & Musen, M.A. (2013) WebProtégé: A Collaborative Ontology Editor and Knowledge Acquisition Tool for the Web. *Semantic Web Journal 4(1)*, 89-99, IOS Press.
44. Vickrey, D., Bronzan, A., Choi, W., Kumar. A., Turner-Maier, J., Wang, A. & Koller, D. (2008). Online Word Games for Semantic Data Collection. In *Conference on Empirical Methods in Natural Language Processing* (pp: 533-542). ACL.
45. Waitelonis, J., Ludwig, N., Knuth, M. & Sack, H. (2011) WhoKnows? - Evaluating Linked Data Heuristics with a Quiz that Cleans Up DBpedia, *International Journal of Interactive Technology and Smart Education, 8(4),* 236-248.
46. Wang, A., Hoang, C.D.V. & Kan, M. Y. (2012). Perspectives on Crowdsourcing Annotations for Natural Language Processing. *Language Resources and Evaluation*, Published online. DOI: 10.1007/s10579-012-9176-1.
47. Weichselbraun, A., Wohlgenannt, G. & Scharl, A. (2010). Refining Non-Taxonomic Relation Labels with External Structured Data to Support Ontology Learning. *Data & Knowledge Engineering, 69(8),* 763-778.
48. Wiggins, A. & Crowston, K. (2011). From Conservation to Crowdsourcing: A Typology of Citizen Science. In *Proceedings of the 44th Hawaii International Conference on System Science (HICSS-44).* IEEE Computer Society.
49. Wohlgenannt, G.,Weichselbraun, A., Scharl, A. & Sabou, M. (2012) Dynamic Integration of Multiple Evidence Sources for Ontology Learning. *Journal of Information and Data Management 3(3)*, 243-254.
50. Wolf, L., Knuth, M., Osterhoff, J. & Sack, H. (2011). RISQ! Renowned Individuals Semantic Quiz: A Jeopardy Like Quiz Game for Ranking Facts. In C. Ghidini (Eds) *Proceedings of the 7th International Conference on Semantic Systems* (pp: 71-78). ACM.