The uComp Protégé Plugin: Crowdsourcing Enabled Ontology Engineering

Florian Hanika [1] florian.hanika@wu.ac.at
Gerhard Wohlgenannt [1] wohlg@ai.wu.ac.at
Marta Sabou[2] marta.sabou@ifs.tuwien.ac.at

[1] Institute f. Information Business, **Vienna Univ. of Economics (WU)**, Austria
[2] **Technical University (TU), Vienna**, Austria

EKAW 2014
November 28, 2014

Introduction - What is this about?

- The **integration** of crowdsourcing into the ontology engineering workflow with the help of a Protégé plugin.
- Certain tasks like validating concept relevance, subClassOf relations, etc. are outsourced to crowd workers.
- Evaluate the impact on quality, cost and time of using crowdsourcing instead of domain experts.

Research Questions

- Which tasks can be crowdsourced? ($\rightarrow$ literature review)
- How to support crowdsourcing enabled ontology engineering? ($\rightarrow$ build a tool)
- Is crowd-based ontology engineering effective and scalable? ($\rightarrow$ evaluation)

Human Computation (HC)

- . . . using human input for tasks/problems that computers cannot solve (yet)
- Different types: **Games with a Purpose, paid crowdsourcing**, and altruism.
- Typical (mechanised labour): paid crowdsourcing to solve micro-tasks, such as:

  `Is every` **`dog`** `a` **`mammal`**`?`

Related Work / Existing Efforts

Crowdsourcing already **widely used** in different areas, for example in

- **NLP community:** GATE Crowdsourcing Plugin [Bontcheva et al, 2014],
- ZenCrowd (entity linking), CrowdMap (ontology alignment)
- Validation of subClassOf relations [Noy et al., 2013]

Why use HC in Ontology Construction / Engineering?

1. Ontology creation typically needs domain experts and is **expensive and cumbersome**
2. Ontology learning often used to **bootstrap** the ontology building process
3. Using the crowd is feasible to further **reduce cost and effort**
4. It has been demonstrated that crowd-workers provide high-quality assessments, especially for general knowledge domains [Noy et al., 2013]
5. Crowdsourcing has been proven useful, but high investments necessary for setting up infrastructure and task $\rightarrow$ **tool support** needed, eg. from inside Protégé

Typical Ontology Engineering (OE) Tasks in Literature

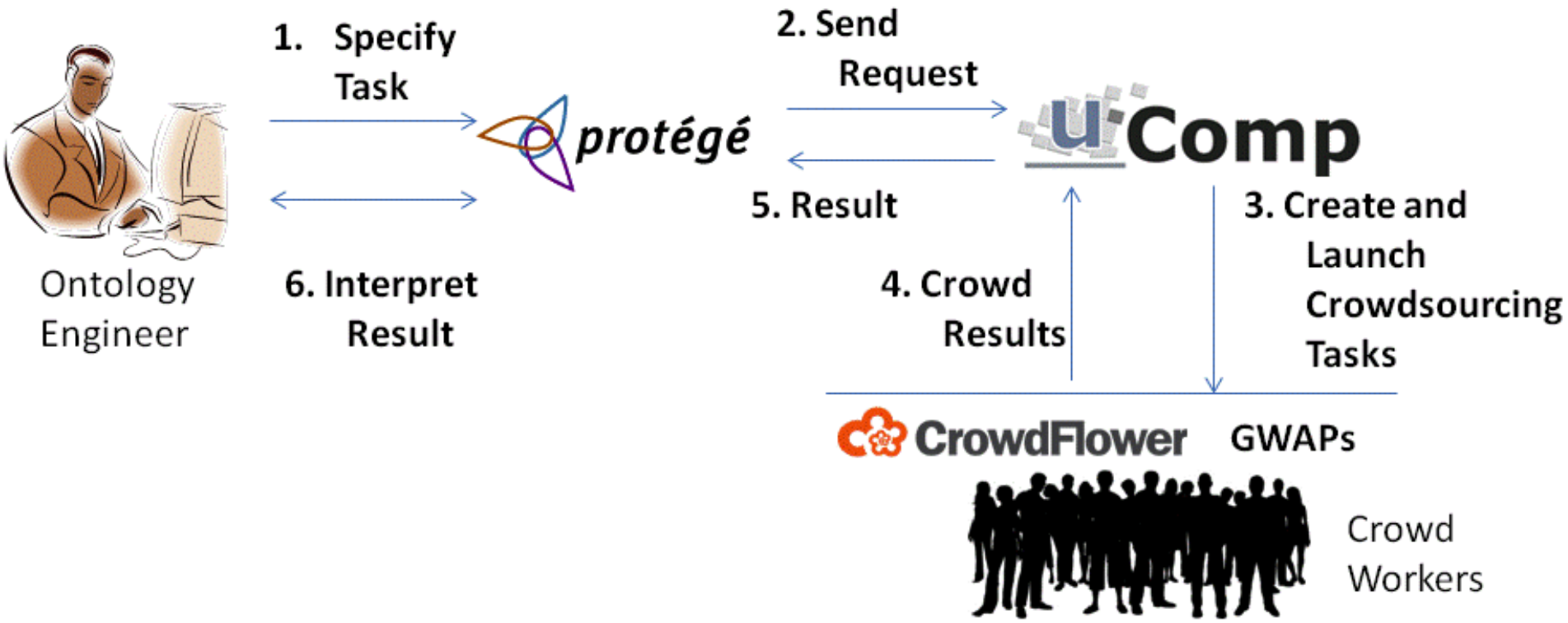
Step 1: which OE tasks?

. . . where crowdsourcing has been applied or appears feasible

- Specification of **Term Relatedness**: are two terms related?
- Verification of **Relation Correctness**
- Specification of **Relation Type**
- Verification of **Domain Relevance**

The Workflow of the Plugin

Figure: Main Stages when using the uComp Plugin



The uComp API

- REST API
- Communication between the Protégé plugin and the crowdsourcing platforms
- **configuration file**
  - **API key:** can be obtained from us
  - **Number of judgements** per unit
  - $-cents paid per judgement

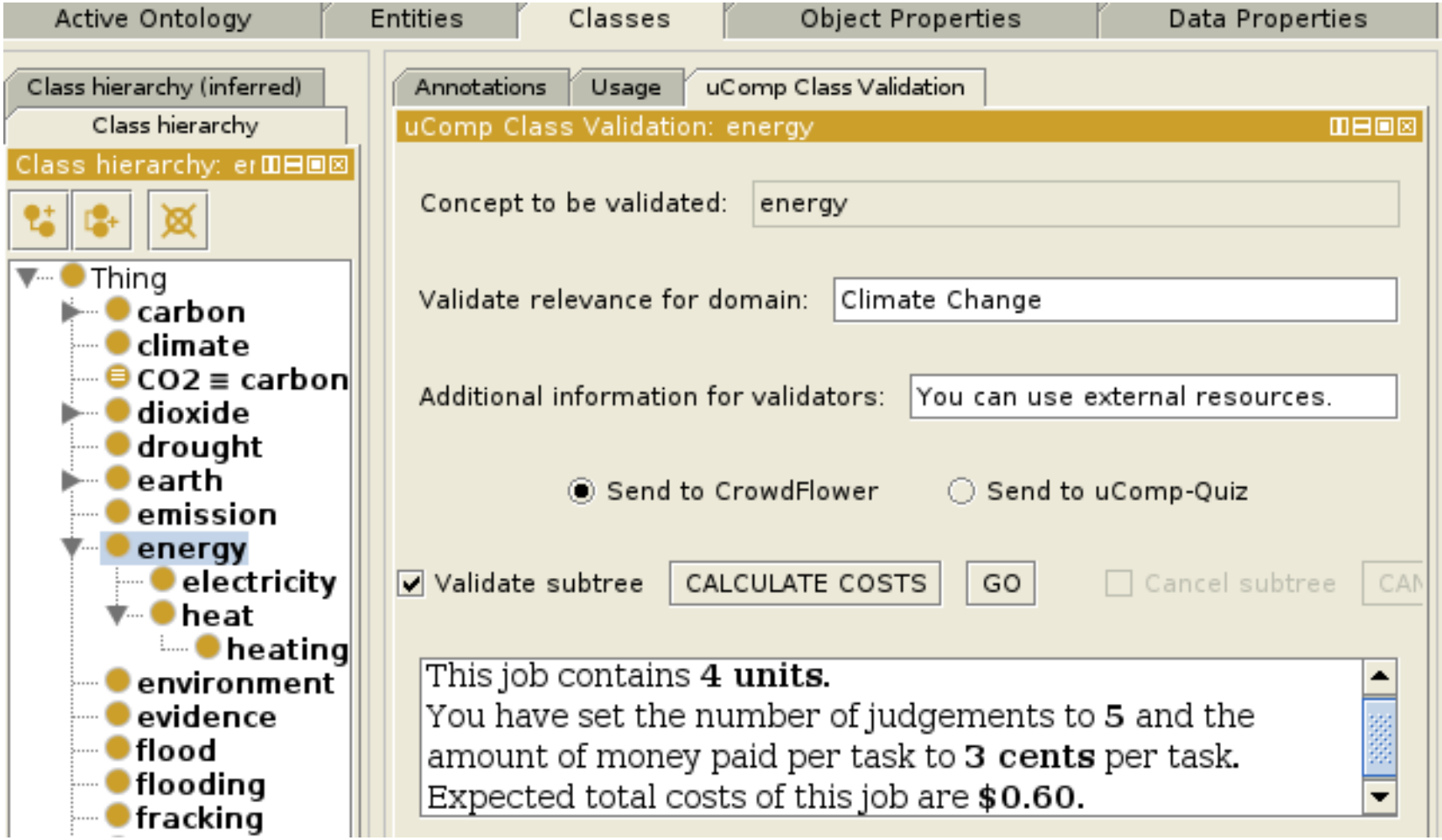
The Protégé Plugin

Step 2: Tool Support

- Plugins usually done as **views** in Protégé
- Accessible for download from inside Protégé (via plugin repository) – named uComp Crowdsourcing Validation plugin

Domain Relevance Verification

The user interface

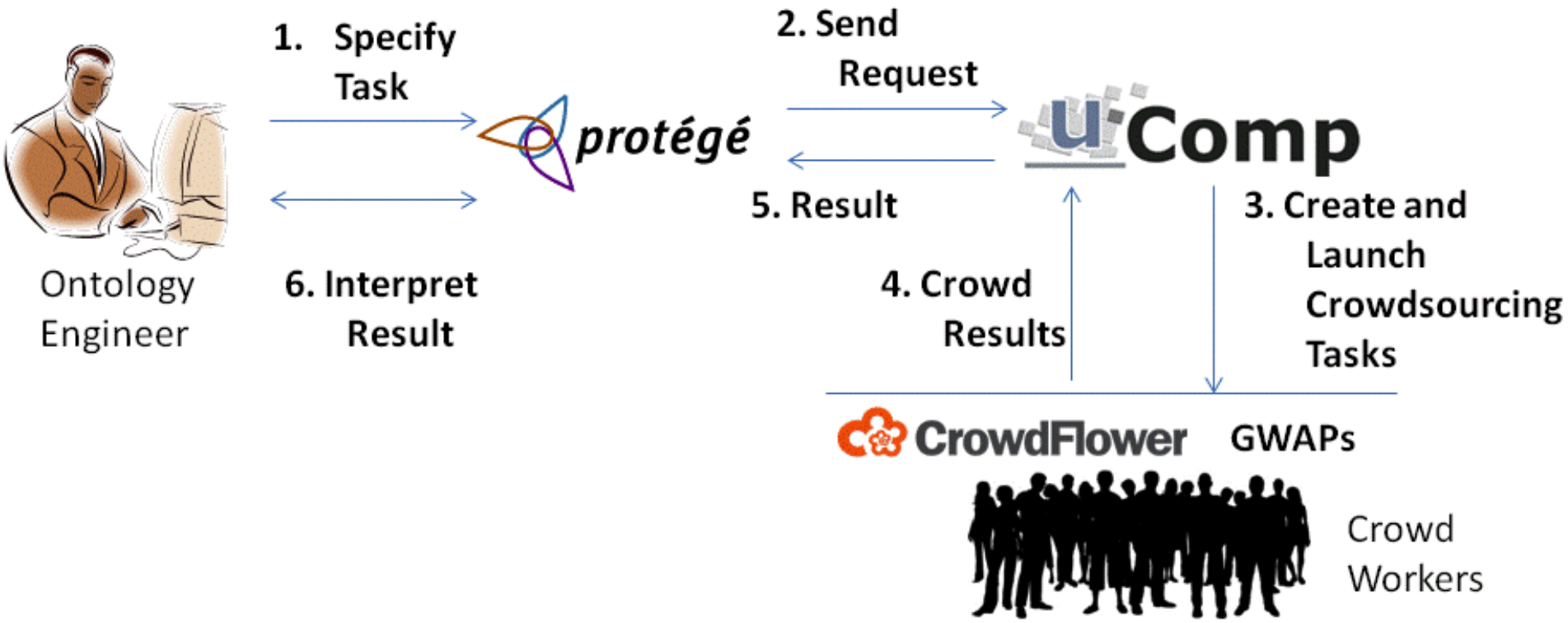


Typical Elements of the UI

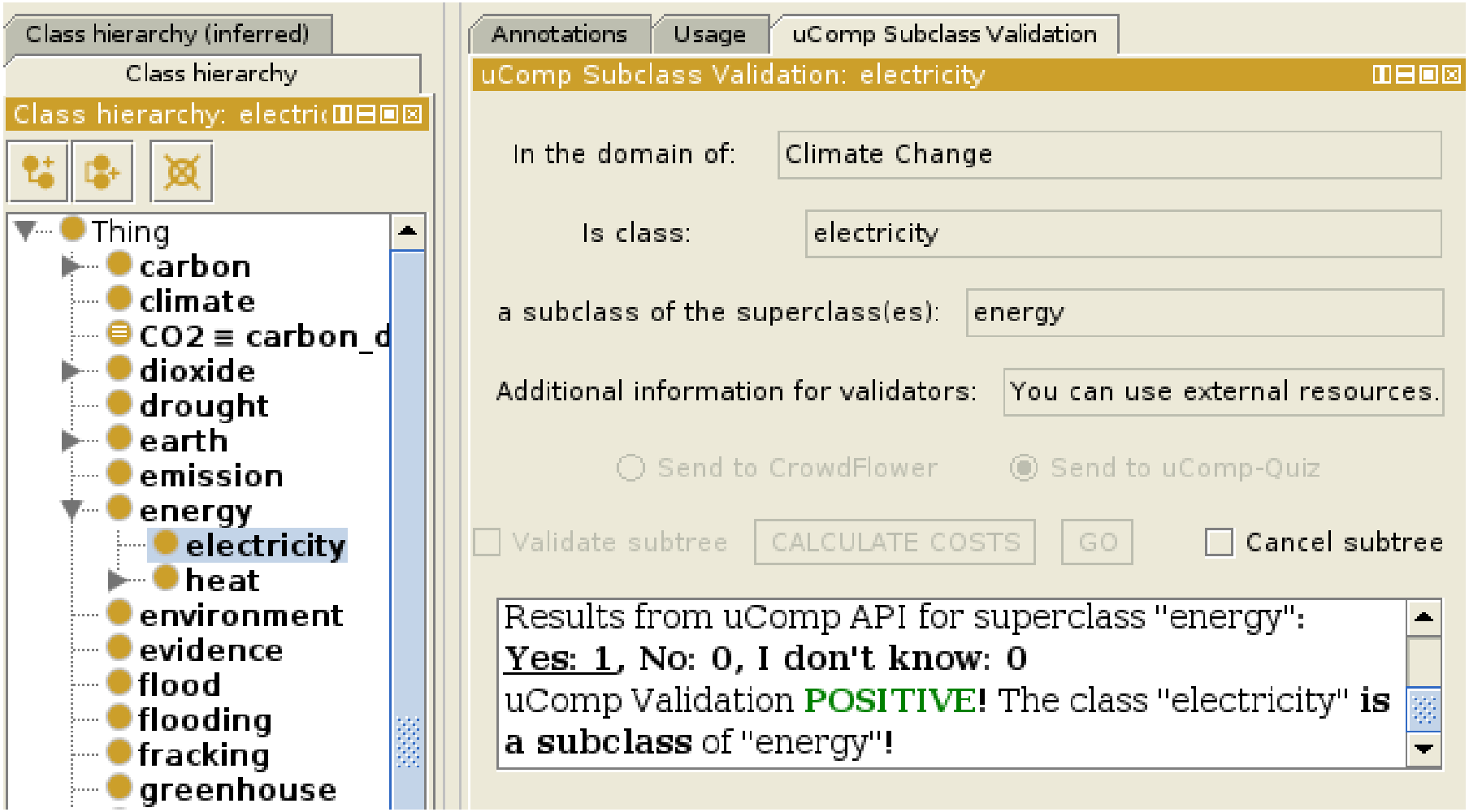
- Task-specific information: eg. concept to be verified
- Generic Information: ontology domain, etc. (pre-filled)
- Task description: Predefined descriptions (for CF users) can be extended here
- Recursive control: also verify all subconcepts?
- CrowdFlower or GWAP?
- Calculate cost

The Workflow

Figure: Main stages when using the uComp plugin

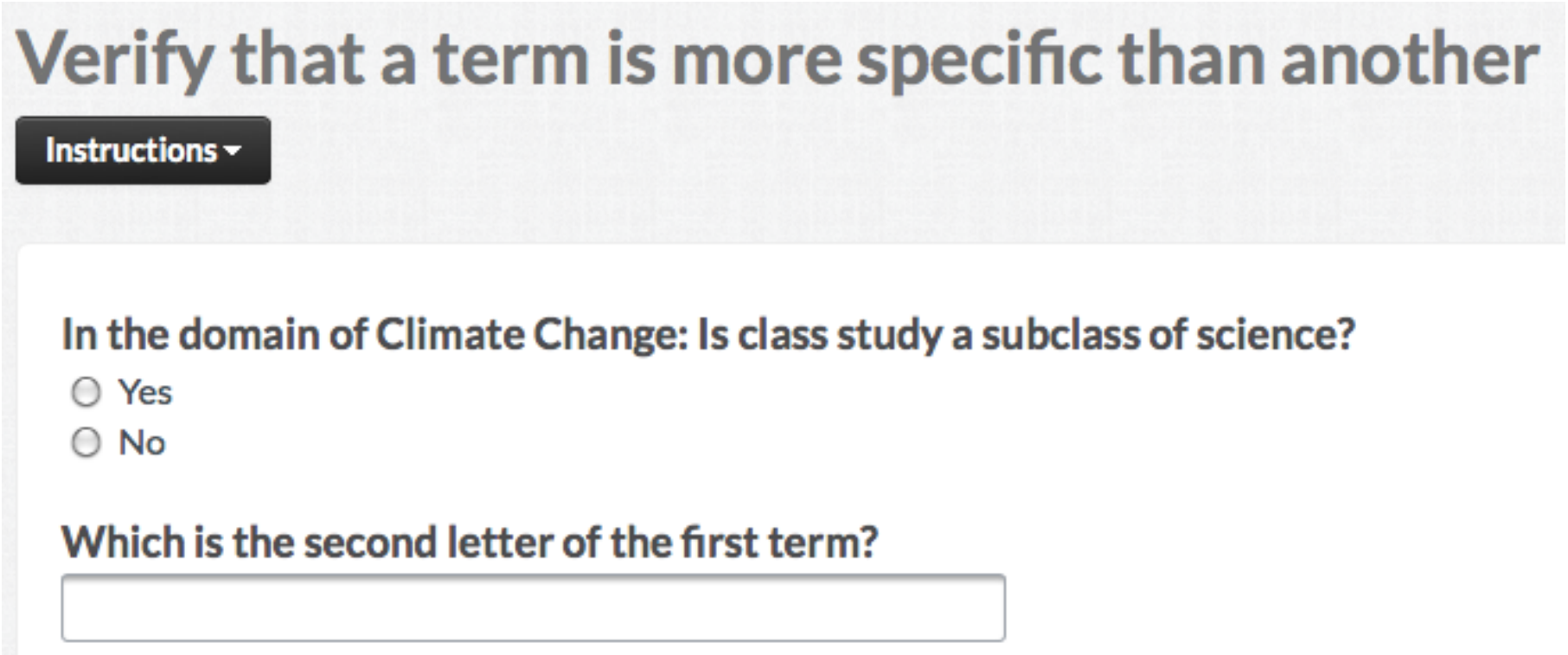


subClassOf Relation Validation

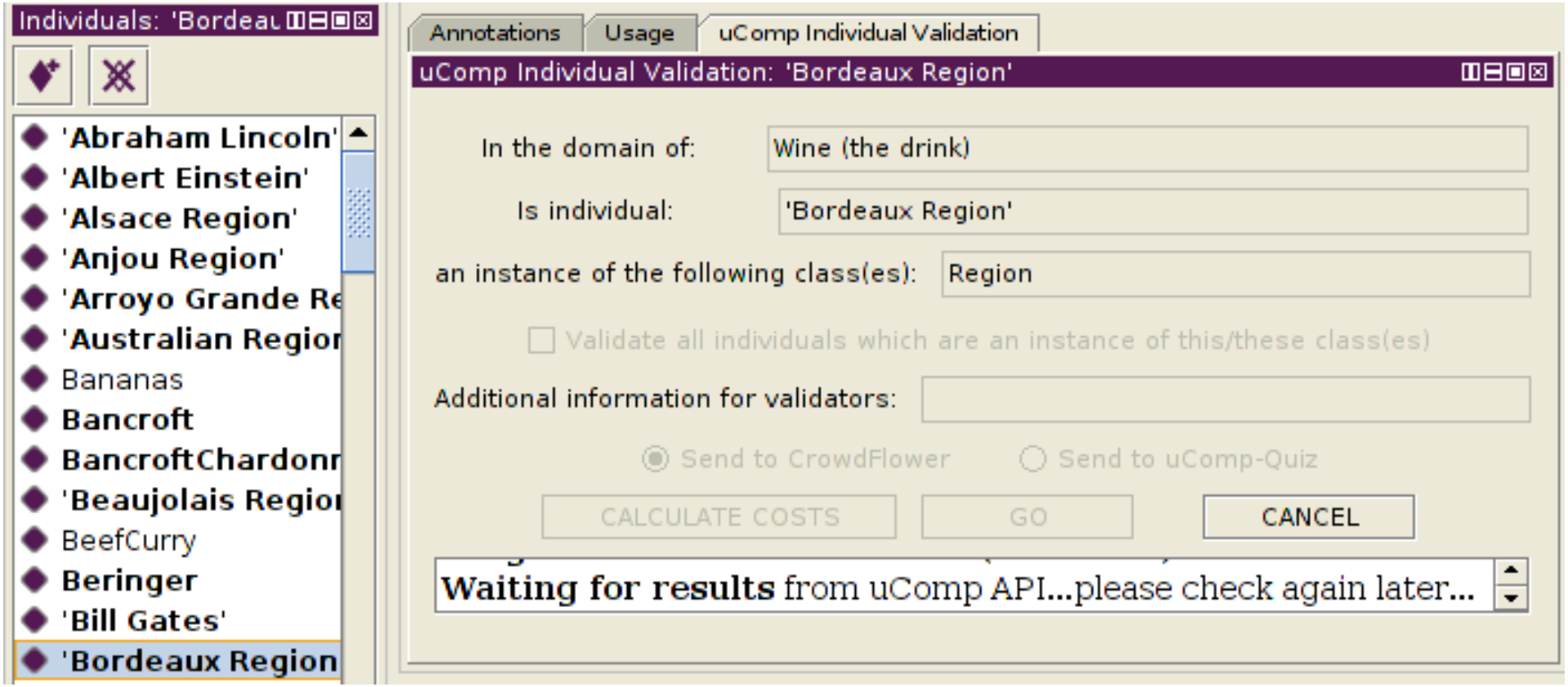


Generated CrowdFlower Job Interface

subClassOf Task



InstanceOf Relation Validation



Setup of Feasibility Evaluation

Step 3: Evaluation

First: Feasibility

- **Domains**: Climate change and finance
- **Ontology size**: ca 100 concepts, ca 40 subClassOf relations per ontology
- **Evaluators**: 8 domain experts (Manual setting), 4 experts using CrowdFlower (Crowd setting) instead

Evaluation Tasks

- **Domain relevance**: For each concept decide if it is relevant for the domain
- **Subsumption**: For each subClassOf relation – verify if it is correct.
- InstanceOf relations

Results – Feasibility – Time and Cost

	Climate Change Ontology	
	Domain Rel	**SubClass Validation**
Time Manual	27.4	23.0
Time Crowd	242	282
Time Saved	0.11	0.08
Cost Manual	11.9	9.9
Cost Crowd	2.1	1.33
Cost Saved	83%	86%

Table : Task duration in minutes and costs (in $)

Results – Feasibility – Agreement

	Climate Change Ontology	
	Concept Relevance	**Subclass Val.**
Manual	0.338 (8)	0.502 (8)
Crowd	0.633 (4)	0.841 (4)
ManualCrowd	0.392 (12)	0.582 (12)

Table : Fleiss' Kappa values of inter-rater agreement per setting and when combining the data of the two settings.

Interpretation: The agreement between the crowd and experts is higher than among experts, possibly because crowdsourcing data is the majority view.

Setup of Scalability Evaluation

- Done for SWJ extended version of the paper
- Assess the plugin's effect on the time, cost and quality for large, real-life and domain specific ontologies
- **Domain**: anatomy (medical) – human.owl ontology which represents the human anatomy part of the NCI thesaurus (from Anatomy track of the Ontology Matching Initiative 10)
- **Ontology size**: 4304 concepts, 3761 subClassOf relations
- **Evaluators**: time and cost estimations for experts, based on wage of a research scientist of $26 per hour
- **Ontology**: we assume it is correct, can use it as a baseline

Introducing Errors

- **Domain relevance check**: we randomly added **1000** concepts from the domains of tennis and climate change
- **Subsumption validation**: we randomly selected 800 leaf pairs of concepts, and swapped them $\rightarrow$ introduce 1600 wrong subClassOf relations

Results – Scalability

	Concept relevance Verification		Subsumption Verification	
	Crowd	**Est. Manual**	**Crowd**	**Est. Manual**
Time (Hours/Days)	19/0.8	19.65/2.5	136/5.6	39.20 (4.9)
Cost ($)	104+26TF	511	155+39TF	1019
Cost (Percentage)	25%	100%	19%	100%
Quality (Accuracy)	0.99	–	0.895	–

Table : Results of the large scale evaluation for the two tasks. Values of the manual approach are estimated based on the results of the feasibility experiments.

Summary of Evaluation / Interpretation

- **Feasibility study**:
  - **Cost:** reduction of 40% to 83% depending on settings used
  - **Quality:** Comparable with that of tasks performed by ontology engineers
- **Large scale / medical domain**:
  - good quality results (accuracy of 89% and 99%)
  - **Completion Time**: similar to domain experts
  - **Cost reduction**: of 75% to 81%

Quality Control in Crowdsourcing (CrowdFlower)

- Provide Gold units – easy questions with predefined answers to filter spammers and bots
- Geographical selection: Eg restrict workers to UK, US, AUS, ...
- Pay reasonably – if feeling screwed, crowdworkers will rush through units
- Only accept high-quality (Level 3) CrowdFlower contributors

Conclusions/Contributions

Summary: Presented the uComp Protégé plugin and its evaluation

Contributions:

1. Identified ontology engineering task suited for crowdsourcing
2. Implemented a Protégé plugin
3. Evaluations show significant cost reduction through crowdsourcing while providing acceptable quality

Future Work

Future work:

1. A clearer methodology and guidelines of where crowdsourcing is helpful and where not.
2. Towards expert sourcing: Investigate engaging the community in a similar fashion.
3. Plugin development: new tasks, improved task monitoring, etc.
4. Further scenarios (like ontology matching), extended usability studies.

Thank you!

Results – InstanceOf Verification

	No. instances	Accuracy	Avg. time (Min)	Avg. Cost ($)
Manual	116	1.0	45.6	19.76
CrowdFl.	116	1.0	120	6.31

Table : Results of the InstanceOf Verification (T2) task.

Results – Relation Type Suggestion

	Climate Change	Tennis
S_Manual	0.536 (5)	0.366 (5)
S_ManualCrowd	0.531 (6)	0.368 (6)

Table : Fleiss' Kappa values of inter-rater agreement for the Specification of Relation Type (T3) task, for two domains.