

Chapter 2

State-of-the-Art: Semantics Acquisition and Crowdsourcing

Abstract In this chapter we review the field of semantics acquisition to provide ground for further discussion on the semantics acquisition games. First, we cover the necessary definitions and review the main “client” approaches for semantics utilization—the information retrieval applications. Then, we move through three major groups of semantics acquisition approaches. The first group constitutes the expert-based approaches: costly, yet often essential for certain tasks such as seeding, setting-up schemas and semantics acquisition output validation. As second, we review the automated approaches: quantitatively effective, yet with questionable quality of output, widely utilized for many tasks such as ontology learning or resource metadata acquisition. Finally, we review the crowd-based approaches, which represent a balance between quality and quantity. They comprise many working schemes, ranging from “explicit” mechanical turking, to “implicit” social tagging applications and of course semantics acquisition games.

2.1 Semantics: Forms and Standards

In general, the term semantics stands for “meaning” and has many interpretations depending on the domain it is used in (e.g., philosophy, linguistics, programming). In the web science field it stands for a formally and explicitly represented meaning of resources (of the Web) readable and “understandable” by machines. At first look, such definition covers only the direct descriptions of resources (such as tags describing an image), however, in practical use, it covers also “not-so-direct”, formally expressed facts, which are also part of meta-information layer above the Web, the domain models (which define the axioms and facts about the world they model). In our work, we use this broader definition. The *web* semantics is then the whole layer of semantics above the web resources. The Semantic Web is its subset that uses specific formalisms, which are widely accepted in the community as semantics representation. The semantics comes in different flavors according to their representations and also

their use. In our work, we recognize two major types: the core semantics (domain models) and resource descriptions (also referred as annotations or resource metadata).

The core of the Semantic Web are the ontologies, which model the *domain knowledge*: the common facts of the world (e.g., “carnivorous animals eat animals”), abstract or concrete concept definitions (e.g., “mammal is an animal”) or constraints and rules of the world (e.g., “mammals have between 2 or 4 legs”). Ontologies are the world abstraction and represent an worldwide agreement on domain’s general rules.

Ontologies that are highly structured and elaborated are often called the “heavy-weight” semantics. Though they provide advanced capabilities, they are also hard to obtain and are almost exclusively built by (costly) human experts. The cheaper and more easy-to-create are the “lightweight” semantics with “lightweight” representations. These include *taxonomies* (hierarchical organizations of entities) and free association networks (which express general relatedness of entities) sometimes referred to as *folksonomies* (in case they were created using a crowdsourcing technique). The semantic structures sometimes use simple terms instead of concepts (also due to ease of creation), which of course, brings drawbacks like semantic ambiguity.

The prevailing standards for ontology representation are RDF (resource description framework) and OWL (web ontology language). The basic elements of ontologies are atomic facts expressed in the form of *triplets* consisting of the source and target entity (*subject* and *object*) connected with a *predicate* (e.g., “dog” (subject)—“is a” (predicate)—“mammal” (object)). Triplets represent various types of relationships among entities (e.g., hierarchy, composition, usage). Each entity or predicate has a textual description, but also URI that identifies it globally. The ontology entities represent *concepts* (classes) or their instances. Each entity can also be decorated with literal properties. Using these basic elements, more complex facts are composed.

As second part of the Semantic Web, *resource descriptions* are the connection between the domain models and web resources. For example, they may denote to which particular concepts the resources are related (e.g., “this article is about bears”) or they provide additional structural information about particular document (e.g., document outline based on semantic properties of individual paragraphs).

Resource annotations are represented either in the knowledge bases (e.g., as RDF triplets) or as a direct part of the web resources itself (e.g., meta-tags of HTML, RDFa). They also vary in terms of what kind of resource (or its part) they annotate. Different kinds of annotations are found in case of texts (the annotation can be related to whole text, paragraph, sentence or even single word), images (spatial information), audio tracks or videos (temporal information).

2.2 Semantics in Use

The web semantics is utilized in various applications. Typically, they are connected to the information retrieval, information space organization, navigation or recommender systems.

Since only a fraction of the Web is covered with descriptive semantics, the applications utilizing them are also limited. Typically, a “semantic application” is dedicated to some specific corpus (e.g., an e-shop search and navigation supported by product descriptions). Heavyweight semantics are not common—applications rely more on lightweight semantics (e.g., product taxonomies).

Today, the query-based search fueled by keyword indexing approaches represents a dominant information retrieval paradigm. A majority of Web users utilize it as their primary way to satisfy their information needs. It also has many drawbacks: expressiveness limits [14], keyword ambiguity [36], invisibility of information space [44], just to mention a few. To solve these issues, some researchers and practitioners suggest a radical change of search paradigm (e.g., exploratory search), others argue for solutions that would not disturb the user who is unwilling to change his keyword search habits (e.g., result re-ranking, query expansion). Though, almost all alternatives somehow rely on semantics.

2.2.1 Query-Based Search

An example of the use of semantics for improving keyword-based search is solving a problem of *term meaning disambiguation* [33]. Searchers utilizing keyword search often encounter problem with homonyms used in the query—their result set gets spoiled with irrelevant (from their perspective) results, because the search terms have multiple meanings. Not all users are able to overcome such issues by themselves.

Researches employed different strategies for solving the search term ambiguity issue. The utilization of semantics plays a vital role in them. Köhler et al. [33] used existing ontologies to index a corpus of websites. They disambiguated terms within the websites using term relationships from the ontologies and were therefore able to infer related concepts, not just keywords. Using the same strategy for search queries, they were able to match relevant search results more accurately.

Another approach to term disambiguation implements the modeling of the searcher. The idea is to track the user’s desires, long term interests or context, represent it in a formal model and measure its relatedness to the potential search results either by query enhancements or result filtering. Comparison is used to re-rank the results, so they satisfy the searcher. Though some researchers [6, 36] attempt to do it on the syntactical (keyword) level, much better results were achieved, if the user and resources were modeled using “heavier” semantics [5].

The search query expansion approaches [4, 9] directly involve searcher’s actions in query disambiguation. Normally, when searcher formulates an ambiguous query, he tries to reformulate it by introducing other keywords, if he can think of any. The point of query-expansion approaches is to aid him with this by recommending a possible search terms to append to the query and to refine the search. Prior to this, a domain model is required to provide relationships of terms in the original query, to other possible search terms.

2.2.2 Exploratory Search

In 2006, Gary Marchionini coined the term *exploratory search* and marked the birth of a new search paradigm, which tried to solve the problems with invisibility of the Web [44]. The invisibility problem is often formulated as the inability of the Web user (searcher) to formulate a search query in a particular domain, because of a lack of familiarity with domain's jargon and structure (it can also be observed in the above case of ambiguous query reformulation).

Marchionini states that not all web search tasks are sufficiently solvable by keyword search, namely learning and information composition tasks. He also states the problems of inexperienced users new to certain domains, who do not know the domain jargon and are thus unable to type in proper keywords [44]. Marchionini suggests the use of alternative search interfaces that visualize some abstraction of the domain that user can afterward navigate and filter the content rather by browsing, then by query formulation. However realization of such interface requires a suitable semantics: both resource descriptions and domain models.

A very basic example of an exploratory search tool is a tag cloud. A relatively widespread technique for result filtering visualizes an information space as a collection of words which characterize its content—the individual resources. The words are usually displayed to the searcher in different sizes reflecting their weight (e.g., importance, frequency) in the corpus. The user is then allowed to review the cloud, learn about the contents of the otherwise invisible corpus. Moreover, he can interact with the cloud by selecting its individual words to filter the result set.

Usually, tag clouds operate over keywords assigned to the resources. However, they could be used with more elaborate semantics too. One example of a search and navigation application exploiting ontologies in this way is *Idea Navigation* [63]. With simply looking textual interface, it directly uses the triplets (instead of tags) to propose possible options of filtering the result set to users. This allows the searcher to learn more about the domain he is reviewing.

A much more elaborate example of exploratory search approach is a faceted browser which heavily relies on semantics. One of such named *Factic* was created by Tvarožek and Bieliková [67, 68]. The faceted browser is an information retrieval application that uses facets (a filter criteria along with their possible values) as the means of formulating of the search queries (rather than by typed-in keywords). Faceted browsers are typical in e-shops where the visitor can filter the products by their parameters (in *Factic*, the facets are even generated automatically based on metadata in ontological repository to which it is attached). The provision of such approach is apparent, the customer can immediately see what is available and he already has a good overview of the whole information space on the e-shop portal, because he sees the category structure and also a list of faceted criteria and their possible values. This, of course, depends on well devised ontology or at least set of homogeneous resource descriptions.

2.2.3 *Creation is the Harder Part*

Though we have not covered other fields where semantics are used like recommender systems [34] or learning frameworks [8], we have illustrated the important role of semantics in today's Web applications. Many of them still await the "critical mass" of existing web semantics in sufficient quality. We have seen that they do not need a particularly rich (heavyweight) semantics to work. Nevertheless, even the lightweight semantics is not available in sufficient scale.

Regardless on how the semantics are used or which type of resource they describe, their creation usually represent the harder part in the job of semantics-based applications development. Many research works (including this) are devoted solely to this task, leaving the semantics use to others. We now continue in description of the existing approaches to semantics creation, which can be split into three categories: manual, automated and crowd-based.

2.3 Manual (Expert-Based) Approaches

Manual approaches for building semantics rely on individual (or small groups of) experts, who create domain models or resource descriptions. Because of the dedicated human work, controlled environment and expertise the experts produce high quality semantics. Their capabilities are, however, limited quantitatively. Employing an expert in a specialized domain can be very costly. Additionally, experts need to be trained to understand the concept, representation and tools for semantics definition (in this task, they are sometimes aided with software tools [28]).

Expert work is essential for certain types of semantics acquisition tasks. In ontology engineering, experts are needed for correct definition of the top layers of the concept (class) hierarchy and system of ontology predicates and constraints. Another example is creation of gold standard datasets or "grand truths" used in evaluation of other semantics acquisition approaches.

The ontology engineering represents a field of study of creating ontologies and covers the spectrum of manual (expert) domain model creation approaches. It covers a variety of methods and methodologies comprising (not only):

- Strategies for defining relevant domain concepts [30].
- Methodologies for definition of relationship schema and ontology axioms [20]—a step which greatly influences the expressiveness of the ontology and the capabilities of automated inference over the ontology.
- Specialized methods for improving collaboration among experts during the ontology building [43, 48].
- What strategies to use in ontology mapping (i.e. interlinking multiple ontologies)—a task which mostly involves seeking for equivalent concepts within different ontologies [32].

- How to clean existing ontologies. Realizing repetitive problems and mistakes occurring during the ontology building, some researchers came up with sets of guidelines for clearing inconsistencies within ontologies [27].

The physical creation of the ontology is then done through software tools like Protégé¹ (WebProtégé [66]) or HOZO [47].

Experts are also employed in resource description acquisition tasks, mostly for commercial purposes. Typical examples are photography, 2D textures or press agency databases. Usually, such corpora are not freely available to public (paid services for professionals, e.g., journalists, designers) and do not follow the description standards of the Semantic Web (i.e. using standards like RDF and linking to global ontologies). However, the principles that are used there are similar to those on of the Semantic Web and such corpora could be eventually (straightforwardly) transformed to meet the Semantic Web standards.

Quantity of delivered semantics is the major disadvantage of expert-based semantics acquisition approaches. However, even manually created knowledge base can grow in size, if it is given enough time and effort. In the *Cyc* project,² a general knowledge base has been constructed for 25 years [38]. The project is being developed for commercial purposes (to be utilized, for example, by expert systems), but it has also been made partially published through the *OpenCyc* release—which demonstrated the impressive scale of 47 thousands of concepts and 306 thousands of facts (triplets or property assignments). Other positive aspect of the *Cyc* is its exhaustiveness: it covers—at least on top levels of abstraction—the whole spectrum of human knowledge. Another advantage is the inclusion of common sense facts that are not present in a written form, simply because they are commonly known (e.g., “You cannot remember events that have not happened yet.”).

The *Cyc* knowledge base is heavyweight: it is highly structured and provides also a reasoning engine to answer even complex logical questions. This, however, comes with a price: the eventual volunteer effort to contribute new facts or even the usage of the knowledge base becomes a difficult task due to its complexity. The *Cyc* critics also note numerous gaps in the ontology: while there is enough concepts the relevant relationships among them often miss [72]. But even with these drawbacks, the *Cyc* knowledge base is usable and expandable and can be a good benchmark for evaluation of automated semantics acquisition approaches.

Another good example of expertly created knowledge base is the *WordNet*³ dictionary [24], used by many web semantics projects. Created and maintained at Princeton university since 1985, it contains English language words organized by *synsets* (according to their synonymic relatedness), parts of speech, lexemes (various textual forms of the same term) and other relationships (e.g., hypernyms, holonyms). In comparison to *Cyc*, *WordNet* is a lightweight corpus. It operates over words, not concepts. Its set of relationships between words is limited and very abstract. Axioms,

¹ <http://protege.stanford.edu/>

² <http://www.cyc.com/>

³ <http://wordnet.princeton.edu/>

constraints or literal attributes absent in its structures. This, naturally, reduces the options of its utilization. On the other hand, its upkeep becomes much more cheaper.

Both Cyc and WordNet are examples of originally “old” (1980) initiatives of domain modeling efforts surviving to this day. They have been created before the birth of the Web. When the idea of Semantic Web emerged, it firstly plead for creation of yet another all-covering (web) world model. However, it soon became apparent that such knowledge base could not be maintained centrally.

This problem was answered with Linked Data initiative. The Linked (Open) Data represent a system of interlinked resources, facts and vocabularies grouped into ontologies, each specialized to a specific domain [11]. Linked Data are, in general, lightweight: their common knowledge representation framework is RDF. This is one of reasons of Linked Data proliferation: the contribution of knowledge to such structure is easier than with heavyweight ontologies. Smaller specialized domain models are easier to maintain. The individual ontologies of the Linked Data overlap which yields a plethora of equivalence relationships between them. Linked Data also incorporated older knowledge bases and reached almost universal recognition in the community as a de-facto central entity of the today’s Semantic Web.

2.4 Automated Approaches

Automated approaches to semantics acquisition rely on extraction of facts out of existing (electronic) human-readable knowledge bases. They have been subjects to many research activities, mainly because they do not rely on cooperation with human contributors which are problematically motivated (for cutting them off, these approaches provide much better scalability). Automated approaches can be seen from several points of view:

- **The source corpora and domain.** The corpus that can be mined can be the whole Web, or it’s subset. It can also be a closed repository of documents (usually related to some domain). Generally, reduction of the input corpus naturally influences the quantity (and also quality) of acquired facts, helps in dealing with the heterogeneity of the resources and brings possibilities to exploit repetitive structures established within the corpus (e.g., reducing corpus to Wikipedia brings the advantages of the infoboxes, which contain structured data).
- **The degree of supervision,** or amount of expert knowledge needed to fuel the process. The typical example of supervised approach is a text mining algorithm, looking for occurrence of certain predefined phrase pattern(s) (e.g., “such as”). On the other side, an unsupervised approach example is the latent semantic analysis of texts (mining frequent term collocations). In general, supervised approaches usually provide better precision, while the unsupervised ones may process more heterogeneous inputs with unexpected situations.
- **The type of job** they do. In ontology building, approaches focus on concept identification, concept instance discovery or on relationship discovery which is

further split by the types of relationship they are able to provide (e.g., non-named, taxonomic or typed). Some approaches are dedicated on relationship naming (which is relevant for example when structuring the lightweight semantics). In resource description acquisition, the approaches can be classified by type of resource they aim to describe: texts, web pages, images, music, videos, etc.

2.4.1 Entity Recognition, Instance Population

When extracting facts from natural language text corpora, the first task for the extraction method is to identify the relevant entities, which will take part as subjects and objects of the triplets. These entities are directly mentioned in the analyzed text. Although, we can imagine that the list of entities (concepts or instances) related to a certain textual resource is possibly wider than the list of “meaningful” lexical units actually present in the text. E.g., there is an article about the American president, without the actual presence of lexical units “American” and “president” in it.

Before entity recognition, the text is usually preprocessed by tokenization and lemmatization (or stemming is used) producing basic morph *terms* (or stems). Then, the stopwords are removed. These are usually supportive words carrying no semantic meaning (e.g., prepositions). The stopwords sometimes contain meaningful words too—in cases they are very frequent in the corpus and thus cannot be used to distinct between resources.

After the preprocessing, the entities are selected, depending on the method. A relatively naive approach is the selection of terms belonging to noun part of speech (which are identified using a dictionary like WordNet). In case of named entity recognition, capitalized terms are selected. Some entity recognition algorithms also solve the possible polysemy (multiple meanings of the same lexemes) of terms, for example by exploiting the existing concept collocation database or thesaurus [46].

Note that the *named entities* require quite different approach for identification. They comprise personal and company names, shortcuts, geographical locations, etc. They are usually not present in the dictionaries. Some approaches for named entity recognition rely on building extensive datasets of such names, which do not have to be necessarily manually created. As example, we can take gazetteer lists constructed by machine learning in the work of Kozareva [35]. A particular problem with named entity recognition is meaning disambiguation (introduced by homonyms), which is being solved by approaches working with term contexts [31].

Apart from “document-driven” approaches that are used for annotating documents the “ontology-driven” approaches focus on populating the domain (or general) ontologies by expanding their hierarchical structure. An example of such approach is the *OntoSyphon* (created by McDowell and Cafarella), which has a purpose of finding instances (or subclasses) for a given ontology class [46]. Their approach, designed to work independently on domain it is used in, takes a domain ontology and text corpus (ultimately—the whole Web) as an input and outputs a ranked list of

candidate instances for a class, also given on input. More detailed, the OntoSyphon works as follows:

1. The textual representation of the input class is retrieved from the ontology.
2. Textual phrases and sentences where the input class is used are retrieved from ontology (if present). Alternatively the ontology neighbors of the input class (parent classes, siblings in hierarchy, other related classes) are retrieved to form artificial phrases.
3. These phrases are used as queries for keyword search engine operating over the given document corpus (here, the whole Web can be easily used).
4. The search engine retrieves a set of documents to be mined. Because of the phrase use, only documents with proper term meanings are retrieved. Without the phrase search, i.e. with only a keyword search with textual representation of the input class, the system might encounter polysemy problems. If, for example, there was a class “sea”, by querying it, we would retrieve documents about sea as a part of ocean, but also about SEA information system. But if the phrase is derived out of the existing ontology (e.g., “sea ship”) and used as a query, much more coherent set of documents with a proper term meaning usage would be retrieved.
5. Finally, using the predefined set of sentence templates (e.g., “A is a B” or “A such as B, C, D”), the OntoSyphon matches the texts of the retrieved documents for expressions of the hierarchical subordination of the named entities, with input class being the superior entity. The other participating entities are afterward written to a instance candidate list.

2.4.2 *Relationship Discovery and Naming*

Another group of automated semantics acquisition approaches orients on the discovery relationships between entities. The entities can be anything from the simple terms to refined ontology concepts. In all cases, textual representations of entities are sought in the textual resources and subsequently their relationships are mined. The factual statements are often contained within the single sentence as subject, object (nouns, adjectives) and predicate (verbs), so many approaches focus on mining the sentences for term relationships [51, 60, 71]. Others try to exploit structures like tables and lists to access the relationship expressed through them [15].

An example of relationship harvesting was presented by Pantel and Pennacchiotti [51]. Their approach implemented a bootstrapping technique, which is, when supplied by few examples, able to harvest quality relationships from the natural language text corpus, even the whole Web. The approach is predicate-oriented: it primarily looks for relationship (predicate) occurrence in the corpus and only afterward, it attaches the subjects and objects to it. The method works as follows:

- At start-up a small set of seed expressions of the same relationship is chosen, e.g., “part of”, “consists of”, “comprises”. Its generic pattern is created to cover variations of the expression, e.g., “X of Y”.

- The bootstrapping technique relies on initial retrieval of large set of potential occurrences of the given seed patterns (which are a phrase stubs). The retrieval is done through a web search engine (or similar engine working over some other corpus).
- With a necessary preprocessing (trimming away the HTML, fragments of texts), the candidate sentences are prepared. Not all of them semantically match the start-up relationship, e.g., “wheel of the car” is correct while “house of representatives” is incorrect relation instance to “part of” relation.
- However, if a certain couple of subject-object (features) is recurring with different seeds, the features are arguably the in the given relationship (in this example, we have of course suppressed the algorithm of feature (entity) recognition).

Another and yet similar “predicate-oriented” approach was presented by Sanchez and Moreno [59, 60] who focused on exploration of non-taxonomic relationships which are insufficiently present in ontologies. It extracts domain-related verbs first and afterward tries to acquire their occurrences in the Web (access through search engine, using verb phrases learned from small domain-related corpus).

There is also an interesting work of Weichselbraun et al., which focuses on labeling (i.e. assigns types or names) of the already existing relationships (also stressing non-taxonomic relationships) [71]. The method mines the corpus of texts looking for co-occurrence of entities coupled in unlabeled relationship and looks up for candidate predicates. The process is, however, supervised by two ontologies: (1) which contains a predefined, finite set of possible relationship labels (domain-related), (2) which contains a taxonomy of all the entities involved in the unlabeled relationships. The purpose of the second ontology is to provide additional constraints that are defined on abstract layers of the ontology and thus have to be valid for lower levels too (which effectively means that not all verbs can be assigned as labels to certain relationships, even if they are found by text mining as candidates).

It is also necessary to mention lightweight semantics acquisition. Typically, latent semantic analysis is used as a “generalized vector space method that uses dimension reduction to generate term correlations” [53]. These correlations or co-occurrences of terms form a network of related terms, if we adopt the premise that if certain terms occurs together often, they are somehow semantically related (although we cannot name the relationship). But even such lightweight semantics are usable (e.g., for query expansion). Moreover unnamed relationships can always be processed by naming approaches and promoted to full triplets.

2.4.3 Automated Multimedia Description Acquisition

Despite their heterogeneous nature in terms of quality, automated metadata acquisition approaches are generally used for annotation of large resource collections. As first major group, we take the image description acquisition approaches.

Many approaches aim to identify semantics relevant to content of static images via identification of visual features. All of these approaches involve some degree of supervision. Duygulu and Barnard [22] employed segmentation of the image and associated identified features within individual segments with words from a large vocabulary. The vocabulary was used afterward to identify the semantics of the image. Their evaluation over Corel 5K dataset yielded 70 % correct prediction. Better results were achieved when a probabilistic model was employed by Lavrenko et al. [37].

Feng et al. [25] proposed enhancement to the segmentation approach, which employed the co-occurrence of terms related to images (e.g., tiger—grass occurring more frequently than tiger—building), which also improved output correctness but was more bound to the training data set of images. Improvements were also achieved when information about global and local features were used together [10].

Various approaches use machine learning for image or image region categorization. Techniques such as SVM [17] or Bayes point machine [16] perform well (precisions over 90 % in Corel 5K dataset), but are limited to a small number of categories and lack of training sets to be used effectively for acquisition of more specific metadata.

Due to its non-textual nature, metadata acquisition for image resources is often performed via analysis of their context (e.g., in the web environment) which may contain text or already annotated resources [52, 69, 70]. The acquisition of the semantics of multimedia content (visual or aural) may also involve OCR or speech recognition approaches [13].

Similarly to images, the raw audio resources are extensive and syntactically complex. Automated acquisition of their semantics is complicated. With images, we are usually satisfied with metadata telling us about physical features in them. The palette of metadata types is wider comprising not only track names, authors, publishers but also lyrics, melody, style, tonality, rhythm, motives or even mood the track evokes on listening. For music information retrieval, the latter group is just as important as the first group. They are used for “querying by example”, which have proliferated next to the standard textual querying [42]. Music metadata are also much more abstract and a potential approach for their acquisition needs to perform sophisticated interpretations of the raw music track.

Many music metadata acquisition approaches involve as a first step a transformation of raw music stream to more symbolic representation, such as musical score or rhythm transcription. An approach of Lu and Hanjalic [41] identifies audio elements (natural semantic sound clusters, e.g., a sequence of chords). Authors point out the similarity of these elements to the words in texts (e.g., a sequence of tones can be understood as a sequence of characters). Thus, the music track can be mined for *keywords*, i.e. the most prominent audio elements. Still, these audio “keywords” cannot be used as normal textual keywords (for textual query formulation). Nevertheless, they provide a basis for effective music track comparison.

A different pre-processing technique was devised by Magistrali et al., who transformed the raw music tracks to an extensive XML and then RDF files. These were then interpreted by rules expertly prepared in an ontology and transformed to more

symbolic representations (still in rdf) [42]. A more supervised, ontology-driven approach, than the “keyword” approach of Lu and Hanjalic, which reminds us of the unsupervised TF-IDF.

The preprocessed audio streams are subjected to further analysis, detecting more complex features and patterns of the music, eventually giving out the desired metadata about their aural characteristics. The unsupervised approaches produce unlabeled features (used mainly in example querying) using mostly statistical process modeling and machine learning [40, 50, 56]. There are also supervised, ontology-driven feature identification approaches [65]. Apart from content-based, also context-based approaches are used [26, 61].

2.5 Crowdsourcing

Crowdsourcing. The term itself was first coined in 2005 by Howe [29]. In 2008, Daren C. Brabham defined it as “an online, distributed problem-solving and production model”. Crowdsourcing (and crowd-based approaches for semantics acquisition) emerged along with the Web 2.0 phenomenon, which enabled masses of Web users to be contributors of the Web content. The crowdsourcing often comprises *human computation* and is focused towards solving of the *human intelligence tasks*—tasks hard or impossible to be solved by computers, but relatively easy for humans. As Quinn and Bederson remind us, these two terms should not be confused [55]. While the “crowdsourcing” primarily designates the distribution of a task to the wide and open mass of people, the “human computation” designates the using of human power for solving of a problem with a computational nature (i.e. a problem that may be solved by computers at some point in the future).

The semantics acquisition involves many tasks performed via crowdsourcing. Users of the Web are time-to-time (and in various contexts) motivated to disclose some descriptive information about web resources they encounter. They comment and rate images or videos, manage their personal content applications, galleries and bookmarks. By collecting these information and tracking user behavior, crowdsourcing techniques produce resource descriptions and even lightweight domain models.

If the crowdsourced semantics originates from the human work, then what differences it have to expert approaches we mentioned earlier? The answer is the different quality assurance mechanisms. While manual approaches rely on an expertise of the individual, the crowd-based approaches the agreement principle: if many, even uninitiated people independently express the same fact, it is probably a truth (e.g., the same photo gets decorated with same tag from multiple users). This allows crowdsourcing to produce relatively precise outputs even if the input is noisy (an individual uninitiated user may produce many untrue suggestions).

The advantage of crowdsourcing approaches against the expert-based approaches is much greater scale of discovered semantics. First, the quantity of potential lay (non-expert) contributors is larger (even when they are used redundantly). On the

other hand, experts are sometimes unavailable. Second, lay contributors are much cheaper or even free (resp. they are not paid for solving a task).

2.5.1 Crowdsourcing Classifications

In general, crowdsourcing comes in many different flavors. It also has very strong overlap with other terms such as human computation, social computing, collective intelligence, crowd computing. Together, they comprise a loosely bounded field and several researchers reflected the lack of abstract, formal models describing it. This resulted into several survey publications, attempting to conceptualize the field with variety of classifications [19, 21, 23, 54, 55].

In his position paper, Erickson classifies the crowdsourcing systems according to distribution of the crowd in time and space [23] (being either at the same time/place or not). This results into four categories of crowdsourcing:

- *Audience based*, when entire crowd participates at the same time and space.
- *Event based*, when the crowd is geographically distributed, but works at the same time on a common goal (e.g., innovation competition).
- *Geocentric*, when the work is done at a particular geographical location by multiple workers in different times (e.g., communal problem reporting).
- *Global*, when the process is bound neither to time nor space.

The typical Crowd-based semantics acquisition approaches (e.g., semantics acquisition games) are found in the latter category (global), as there is usually no need to bound them to specific time or place (though as a significant exception, various geocentric applications for collecting metadata on points of interest, e.g., FourSquare, should be mentioned).

Doan et al. defines nine dimensions according to which the crowdsourcing applications could be considered [21]. We look at six of these dimensions interesting through prism of crowd-based semantics acquisition:

- *What type of target problem is being solved (e.g., labeling images, building a knowledge base, rating movies)?*
- *What is the nature of collaboration?* Authors identify two major groups of approaches: *explicit* (where users explicitly collaborate to create a useful artifact as their primary objective) and *implicit* (where users solve a target problem as a side effect of another activity). The semantics acquisition tasks fall in both categories. The explicit approaches include item rating or knowledge base (ontology) building, the implicit comprise for example, image tagging through crowdsourcing games.
- *How does the application recruit and retain new workers?* This perspective brings up the question of incentives to the workers. Some semantics acquisition applications are useful for the user himself (e.g., tagging websites in bookmark portal), some rely on volunteers (e.g., contribution to knowledge bases like ConceptNet), some motivate by entertainment (e.g., crowdsourcing games).

- *What do users do in the process (how they solve the tasks)?* What technique is used by users to contribute (e.g., tagging, rating, reviewing)? How cognitively- or skill-demanding? Here, applications that ask simpler questions retrieve more answers from more workers but may also demand more validation (e.g., when workers validate existing metadata by dichotomic yes-no options). Is the semantics (being retrieved) a common sense knowledge or a specialized domain knowledge (which is obviously harder to obtain due to smaller pool of potential contributors)?
- *How are the partial results combined?* And how is the problem decomposed prior to that? In semantics acquisition, many approaches tend to collect atomic pieces of information (tags, triplets) which then (automatically) constitute more complex structures. A contrast to this are, for example, contributors to Wikipedia, that compose complex structures (texts) themselves and where the contribution combining is a demanding human intelligence task itself.
- *How is the output evaluated?* Common for semantics acquisition is redundant task solving and collaborative filtering—a technique possible mainly due to the “atomicity” of the acquired information. Apart from this, however, other techniques exist, such as rating of contributions created by other users, post hoc cleaning by domain experts or detection of suspicious behavior of workers (and thus, malicious contributions).

At approximately same time as Doan et al., Quinn and Bederson offered a different conceptualization and design space of the “combined” fields. Although they focused primarily on the human computation (rather than crowdsourcing) [55] and offered a classification of human computation approaches, their classification dimensions strongly refer to the crowdsourcing too. For each dimension, Quinn and Bederson also name several values, i.e. typical design patterns or features utilized by human computation systems. At the same time, they declare the list as open and waiting to be filled with new alternatives.

- *Motivation.* What motivates people to contribute? This dimension is directly mappable to Doan’s “recruitment and retention”. As major incentive forms, Quinn and Bederson identifies pay, reputation, altruism, enjoyment and “implicit” work (covered by Doan’s “nature of collaboration”).
- *Quality control.* Another “recurring” dimension of Doan’s: “output evaluation”. Though, Quinn and Bederson offer a finer-grained set of patterns. Many of them represent some kind of redundant task solving, common for semantics acquisition: output and input agreement (a reference to work of Luis von Ahn’s semantics acquisition games [1]), redundancy, statistical filtering.
- *Aggregation.* Describes how are the individual worker contributions combined. A dimension directly mappable to Doan’s “partial result combination”. Authors identify a variety of approaches, including iterative improvement (or validation) of existing artifacts, searching for positive cases (e.g., visual scanning of large set of satellite images for evidence) or evaluation of fenotypes of genetic algorithms. More characteristic for semantics acquisition however, is simple collecting of partial contributions into a larger structure (e.g., atomic facts into an ontology),

feeding training data to machine learning approaches (e.g., training set of images with metadata) or using no aggregation at all.

- *Human skill.* What type of cognitive activity are the workers performing? Authors mention visual recognition, language understanding and human communication. The visual recognition together with aural recognition is often a case of multimedia metadata acquisition approaches. As another human skill category, we recognize the application of “common sense” which is a subject of several knowledge acquisition projects [39]. We see two counterparts to this dimension in Doan’s work: the “target problem (type)” and “how do workers solve the task” (what tools or techniques they use). In both cases however, Doan et al. focus on “outer” characteristics of the job, whereas Quinn and Bederson focus on mind skills themselves. An attempt to categorize human skills used in human computation was also made later work by Parshotam [54], who identifies them as human perception (sensing), cognition, knowledge, common sense, visual processing, anomaly detection or context identification.
- *Process order.* For this dimension, authors identify three roles found in each human computation system: the requester, worker and computer. Then, several classes of systems based on order of work of these roles are presented. Sometimes, the computational task is firstly attempted by a computer and then corrected or complemented by a human, e.g., computer-worker-requester for ReCAPTCHA.⁴ In other cases, the human contribution precedes the computer processing, e.g., a semantics acquisition game Peekaboom, where players identify visual objects by circular regions in the images which are further automatically folded to form true (i.e. non-circular) boundaries of these objects [2]. For semantics acquisition, both cases are common. Moreover, the role of computer processing (either prior or posterior), not only for mediation is often essential to handle the quantity of tasks (high even for a crowd processing).
- *Task-request cardinality.* How many workers are necessary to finish one task?

The authors encourage to further experimentation with the classification by combining various dimensions and their values to imagine new systems.

Based on the literature review, the (1) role of incentives (motivation) and (2) quality control receive most of the attention of researchers in crowdsourcing and human computation ([19, 57, 64] resp. [3, 19, 45, 73]).

2.5.2 Mechanical Turk

As a demonstration and a single most renown product (and at the same time, an approach) of the crowdsourcing the Wikipedia is often presented. A much more characteristic to the crowdsourcing principles however, is the *Amazon Mechanical Turk*.⁵ It is a generic platform for controlled crowdsourcing, where companies or

⁴ <http://www.google.com/reCAPTCHA>

⁵ <https://www.mturk.com>

individuals (task owners) submit tasks they need to solve by a crowd (e.g., annotation of image collection). When the tasks are submitted the Turk organizes contributors to solve them, following the instructions given by task owner (e.g., how many times a particular task instance should be solved, what criteria a contributor must fulfill).

A significant feature of the Mechanical Turk are micro payments to contributors (usually units of cents) for each task solved. Micro payments represent important motivation for contributors to participate in the process. It is sometimes secondary to other, primary motivation, but is necessary. A good example of this is participation in a crowd-based scientific experiment evaluation (e.g., validation of resource metadata): the contributor is sympathetic to the cause, but the definitive incentive to join the process is the money (although small) he receives for the job [58].

Apart from this model, there are also crowdsourcing approaches that employ contributors *without* the need of motivating them by monetary values. Almost exclusively, the semantics are then only a side-product of the user activity, which primarily focused on their needs (e.g., social bookmarks, comments). Sometimes, users do not even know they are contributing to some knowledge base. Due to these facts, possible kinds of semantics we can collect via crowdsourcing is limited with types of activities the users usually do on the Web (although there is always a possibility to attract their attention to some new activity). For example, common users of the Web upload and annotate (textually with tags) images. They do this, because they want to have them organized, always available and shareable to friends (e.g., Flickr⁶ image gallery), not because they should be annotated. In social networks like *Facebook*, users locate exact position of persons in the images. However, we can hardly expect them being motivated to locate non-living things (which also deserve such annotations).

2.5.3 *Delicious*

A typical case of semantics acquisition via crowdsourcing is the bookmarking portal Delicious.⁷ Here, users submit URLs they want to visit later or simply have them at hand for some reason, similarly to web browser bookmarks. Here, however, they have them online so they do not have to create them repeatedly on different workstations and moreover, they decorate them with tags (the submission procedure requests the user to provide some tags). Using the tags, the URLs can be easily filtered and even large set of bookmarks are relatively easy to browse (e.g., by using tag clouds).

From the Semantic Web perspective, the Delicious users do two useful things:

1. They **decorate URLs**, i.e. web resources, *with tags* and annotate them. Unfortunately, they do it with respect to themselves, i.e. they write tags which meaning they understand, but this meaning can be proprietary to them only and therefore confusing or inaccurate for the rest of the world. For example, someone bookmarks the Wikipedia page about grizzly but decorates it with tag “55 km/hr”,

⁶ <http://www.flickr.com>

⁷ <http://delicious.org>

which for the user represents the speed of the bear and makes perfect sense since he is collecting articles to create ranking of animal speed. But from the universal point of view, this tag represents only a marginal part of the article. Another drawback of user tagging is the relative generality of the used words (causing overload of the tags lowering their distinctive ability) and often use of sentiment or even irony with no semantic feedback on the content of the URL-identified resources (e.g., “funny”, “gorgeous”, “stunning”) [49].

2. **Generate tag collocations.** When a user decorates a resource with more than one tag, he expresses the, yet unspecified, relatedness of these tags. These collocations are afterward used to create a lightweight tag folksonomy [49].

Even with all drawbacks, the user tagging is a potent source of lightweight semantics and annotations. Using the agreement principle and dictionary, more clean metadata can be obtained and utilized like, for instance, in the project *Treelicious* which combines the delicious folksonomy with *WordNet* hierarchy to create a navigable tag structure.

The associations between term relationships in folksonomies like delicious are unknown, meaning that we do not know, what does the relationship means. Some researchers build upon the folksonomies like Delicious and leverage their term relationships by identifying their specific types. In their work, Barla and Bielíková [7] extract hierarchical relationship between terms with syntactical analysis of the term graph structure.

2.5.4 Wikipedia and DBpedia

The DBpedia is an example of semantic building approach that mixes all manual, crowdsourcing and automated approaches. It is closely related to world largest and collaboratively created encyclopedia, the Wikipedia. Although the Wikipedia’s articles can be seen as being created manually by pseudo-experts, the large number and the collaboration of the contributors pushes it rather to the crowdsourcing category of knowledge acquisition. While Wikipedia is primarily made readable to human users, the DBpedia transcripts the knowledge contained in Wikipedia to a more “machine-friendly” ontology, using RDF and OWL standards.

The basis of the DBpedia content is created automatically: Each article of Wikipedia becomes a concept in DBpedia. Using various algorithms, the texts and links of Wikipedia are mined to create relationships and assign properties to the concepts. These include [12]:

- Extraction of labels from titles and link named entities.
- Abstract extraction from original article texts.
- Article categories.
- Geo-coordinates.
- Properties through infoboxes. Infoboxes are structured information attached to some articles, consisting of properties and their values. The infoboxes of articles

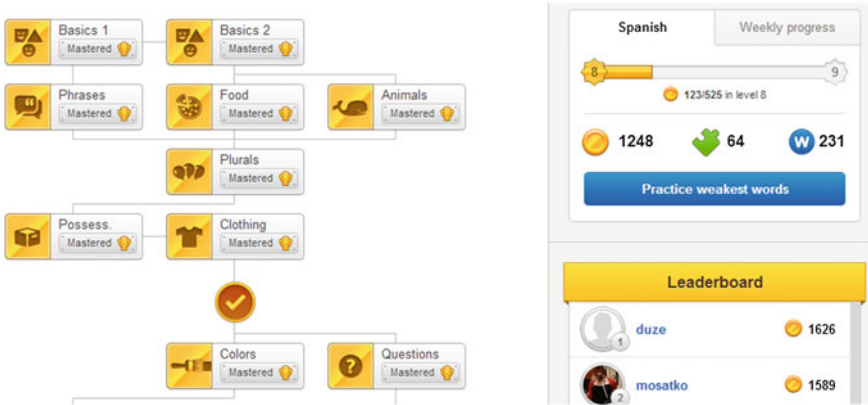


Fig. 2.1 Example of various Duolingo achievements, constantly reminding student of his progress and inviting for more activity

of same category follow the same structure, so it is relevant to use their properties as attributes of the category (a superclass) in the DBpedia (e.g., articles about kings of France are decorated with reign time span and prime minister attributes).

The DBpedia also pays respect to other existing global (semantic) data resources (like FOAF ontology). Unfortunately the automatically extracted facts of DBpedia are still somewhat sparse. Although the ontology contains a solid concept hierarchy, which originates from the manually created and refined Wikipedia classification system, it lacks relevant non-taxonomic relationships (e.g., composition, interaction).

2.5.5 Duolingo: A “Gamified” Crowdsourcing

One of the incentives used to motivate workers to participate in a crowdsourcing process is the gamification: an introduction of game elements (e.g., leaderboards, badges, achievements) into activities that are not games themselves. By its definition, gamification covers any working activity (not just crowdsourcing), with crowdsourcing however, it is with good symbiosis: the small tasks allow fluent rise of player’s “achievement” levels, constantly reminding of his progress. A good example of fusing gamification and crowdsourcing is a language learning portal called Duolingo⁸ created by Luis von Ahn. In Duolingo, a student may learn a new language from scratch, using interactive exercises that automatically evaluate his written or even spoken answers. For this, student receives various achievements and badges (see Fig. 2.1)—he is constantly reminded of his progress.

⁸ <https://www.duolingo.com>

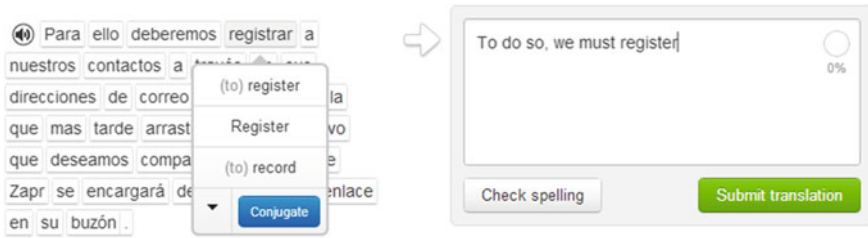


Fig. 2.2 Example of Duolingo's interface for translating real Web. The application aid the student with translation of individual work, making him practice with vocabulary and able to translate even complex sentences

But Duolingo has also its crowdsourcing part: as a practice, its students can translate real sentences from web pages written in the foreign language to English. For this, a convenient interface (see Fig. 2.2) is provided. The student may also review and rate other translations. For these activities, he is also rewarded with achievement points. Using a redundancy principle and a wisdom of the crowd, Duolingo is able to output very accurate translations of a real web pages using only a group of lay translators that are actually *just learning* the language.

2.5.6 Crowdsourcing Games

The gamification aims to solve the problem of limited motivation to participate in a crowdsourcing by using game-like elements. As such, it may be perceived as a springboard towards “full” *crowdsourcing games*—a part of the crowdsourcing approach family. These games emerged as an alternative to solving computational problems, hard or impossible to be solved by machine computation (which includes acquiring semantic structures), via aggregation of knowledge provided by many non-expert users (e.g., for image annotation) [62]. Crowdsourcing games transform problems into games that motivate players to solve them via fun and thus eliminate the need to pay them. As many game instances can be played simultaneously, they are suitable for larger scale problems divisible into smaller tasks. Compared to other crowdsourcing techniques, the knowledge gained in crowdsourcing games is not just a by-product of another user activity (e.g., annotating web resources for personal use), but the *primary* objective, so their design is tuned to maximize that ability. The crowdsourcing games are discussed in detail in the Chap. 3.

2.6 Discussion

To sum up, there is a variety of approaches for building web semantics ranging from manual through crowdsourcing to automated ones. Semantics discovery approaches are evaluated with respect to quantity (number of instances retrieved,

number of document covered, universal versus domain specific applicability) and quality (tolerance to bias and errors, structural degree) of information and knowledge they are able to acquire. Using these perspectives, we observe a generally high quality and low quantity results of expert based approaches that are bound to the limited manpower. On the other hand, automated approaches deliver semantics in high quantities but with unsure quality, since they are prone to unusual situations sourcing from the heterogeneity of spaces they aim to cover. The crowd-based approaches are somewhere in between, operating with numerous, yet lay mass of human contributors. They have potential for both quality and quantity, but are limited by specificity of the task they aim to fulfill. They also need to motivate the contributors the right way, which is also limiting. These (but not only these) issues make the field of crowdsourcing a target for researchers.

Some researchers argue there is no other way to create accurate domain models and annotations, than to utilize manpower, others argue that virtually any piece of knowledge is already on the Web, probably with great redundancy and it is only a matter of developing of the ultimate harvesting algorithm to collect it [18].

For now, the best way toward acquisition of semantics lie in combining approach families together to exploit strong points and neutralize weaknesses. As an example of approach chaining, we can imagine a ontology engineering project where experts firstly set top layers of the taxonomy within the ontology, set up the axioms and entity and relationship types and seed the examples. After this, an automated method is deployed over the corresponding text resource corpus and extracts entities and relationships according to patterns (previously set by expert). Lastly, the crowd comes in to validate the acquired entities and relationships using a simple true/false question answering interface. As another example of symbiosis, we can consider a crowd that prepares image tags for images prior to the automated classifier training.

Considering this, we come to two possible roles of the crowd: semantics creation or semantics validation. Whether the crowd is supposed to carry out first or the latter, greatly influences the options the method designer has. Naturally, a “validation” crowdsourcing always depends on an existing metadata set it aims to improve. On the other hand it has a great advantage regarding the design of the contributor’s interface with the crowdsourcing platform: validating something is in general more ergonomic than creating (both syntactically and semantically). In the context of the first example, a dichotomous question answering about the validity of a typed relationship between two terms is syntactically easier than selecting the type from a long list. This somewhat advocates the use of crowdsourcing for semantic validation rather creation, especially if the automated method that creates the metadata is able to state its confidence (support) about its output, limiting the metadata set that needs to be validated to only “unsure” cases.

The type of the resource for which the semantics is created also indicates the potential outcome of the acquisition method. For structured and unstructured texts, automated approaches function better if only lightweight structures are demanded (e.g., keywords), whereas experts or crowds are needed, if the semantics (especially domain models) is required on a higher quality grade. With multimedia, the human work is even more demanded in semantics creation. For our research presented in

this book, we identified a particular domain of personal image metadata creation, in which all families of approaches are nowadays shorthanded. The personal multimedia creation cannot be subjected to either automated means or crowds, because neither of these possesses the specific knowledge (e.g., awareness about person names or specific places).

As another issue, we address in this book, we identified the upkeep of the semantics. Nowadays, researchers are primarily focused on semantics creation and only little attention is given to already created semantics. Yet, this existing corpora must be constantly reviewed, validated and updated. Many times, metadata are temporal “by definition” or are invalidated by the change of the underlying resource. The metadata may also be wrong from the moment they are created (after all, the automated and crowd-based method are sometimes prone to errors). All of these “effects” may render a metadata corpus partially invalid and a needed subject to cleanup (removing incorrect or invalid facts) and a potential “renovation” of semantics (creating new, correct facts to as substitute the removed). In this work, we chose the domain of music metadata corpora, created by human taggers, as a candidate for metadata cleanup (realized through crowdsourcing games).

References

1. von Ahn, L., Dabbish, L.: Designing games with a purpose. *Commun. ACM* **51**(8), 58–67 (2008)
2. von Ahn, L., Liu, R., Blum, M.: Peekaboom: a game for locating objects in images. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '06*, pp. 55–64. ACM, New York (2006) NULL
3. Baba, Y., Kashima, H.: Statistical quality estimation for general crowdsourcing tasks. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '13*, pp. 554–562. ACM, New York (2013)
4. Bai, J., Song, D., Bruza, P., Nie, J.Y., Cao, G.: Query expansion using term relationships in language models for information retrieval. In: *Proceedings of the 14th ACM International Conference on Information and Knowledge Management, CIKM '05*, pp. 688–695. ACM, New York (2005)
5. Barathi, M.: Context disambiguation based semantic web search for effective information retrieval. *J. Comput. Sci.* **7**(4), 548–553 (2011)
6. Barla, M.: Towards social-based user modeling and personalization. *Inf. Sci. Technol. Bull. ACM Slovakia* **3**(1), 52–60 (2011)
7. Barla, M., Bielíková, M.: On deriving tagsonomies: keyword relations coming from crowd. In: *Proceedings of the 1st International Conference on Computational Collective Intelligence, Semantic Web, Social Networks and Multiagent Systems, ICCCI '09*, pp. 309–320. Springer, Berlin, Heidelberg (2009)
8. Barla, M., Bielíková, M., Ezzeddinne, A.B., Kramár, T., Šimko, M., Vozár, O.: On the impact of adaptive test question selection for learning efficiency. *Comput. Educ.* **55**(2), 846–857 (2010)
9. Bhogal, J., Macfarlane, A., Smith, P.: A review of ontology based query expansion. *Inf. Process. Manage.* **43**(4), 866–886 (2007)
10. Bielíková, M., Kuric, E.: Automatic image annotation using global and local features. In: *Proceedings of the 2011 Sixth International Workshop on Semantic Media Adaptation and Personalization. SMAP '11*, pp. 33–38. IEEE Computer Society, Washington (2011)

11. Bizer, C., Heath, T., Berners-Lee, T.: Linked data—the story so far. *Int. J. Semant. Web Inf. Syst.* **5**(3), 1–22 (2009)
12. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia—a crystallization point for the web of data. *Web Semant.* **7**, 154–165 (2009)
13. Bolettieri, P., Falchi, F., Gennaro, C., Rabitti, F.: Automatic metadata extraction and indexing for reusing e-learning multimedia objects. In: *Workshop on Multimedia Information Retrieval on The Many Faces of Multimedia Semantics. MS '07*, pp. 21–28. ACM, New York (2007)
14. Botev, C., Amer-Yahia, S., Shanmugasundaram, J.: Expressiveness and performance of full-text search languages. In: *Proceedings of the 10th International Conference on Advances in Database Technology. EDBT'06*, pp. 349–367. Springer, Berlin, Heidelberg (2006)
15. Buitelaar, P., Cimiano, P., Frank, A., Hartung, M., Racioppa, S.: Ontology-based information extraction and integration from heterogeneous data sources. *Int. J. Hum Comput Stud.* **66**(11), 759–788 (2008)
16. Chang, E., Goh, K., Sychay, G., Wu, G.: Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Trans. Cir. and Sys. Video Technol.* **13**(1), 26–38 (2003)
17. Cusano, C., Ciocca, G., Schettini, R.: Image annotation using SVM. *Proc. SPIE* **5304**, 330–338 (2004)
18. Dalvi, N., Kumar, R., Pang, B., Ramakrishnan, R., Tomkins, A., Bohannon, P., Keerthi, S., Merugu, S.: A web of concepts. In: *Proceedings of the Twenty-Eighth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 1–12. ACM (2009)
19. Das, R., Vukovic, M.: Emerging theories and models of human computation systems: a brief survey. In: *Proceedings of the 2nd International Workshop on Ubiquitous Crowdsourcing, Ubi-Crowd '11*, pp. 1–4. ACM, New York (2011)
20. Di Maio, P.: 'Just enough' ontology engineering. In: *Proceedings of the International Conference on Web Intelligence, Mining and Semantics, WIMS '11*, pp. 8:1–8:10. ACM, New York (2011)
21. Doan, A., Ramakrishnan, R., Halevy, A.Y.: Crowdsourcing systems on the world-wide web. *Commun. ACM* **54**(4), 86–96 (2011)
22. Duygulu, P., Barnard, K., Freitas, J.F.G.d., Forsyth, D.A.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: *Proceedings of the 7th European Conference on Computer Vision-Part IV. ECCV '02*, pp. 97–112. Springer, London (2002)
23. Erickson, T.: Some thoughts on a framework for crowdsourcing. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI'11. A Position Paper for the CHI 2011 Workshop on Crowdsourcing and Human Computation*. ACM, New York (2011)
24. Fellbaum, C. (ed.): *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA (1998)
25. Feng, S.L., Manmatha, R., Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation. In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR'04*, pp. 1002–1009. IEEE Computer Society, Washington (2004)
26. Ferrara, A., Ludovico, L.A., Montanelli, S., Castano, S., Haus, G.: A semantic web ontology for context-based classification and retrieval of music resources. *ACM Trans. Multimedia Comput. Commun. Appl.* **2**(3), 177–198 (2006)
27. Guarino, N., Welty, C.: Evaluating ontological decisions with ontoclean. *Commun. ACM* **45**(2), 61–65 (2002)
28. Gulla, J.A., Sugumaran, V.: An interactive ontology learning workbench for non-experts. In: *Proceedings of the 2nd International Workshop on Ontologies and Information Systems for the Semantic Web. ONISW '08*, pp. 9–16. ACM, New York (2008)
29. Howe, J.: The rise of crowdsourcing. *Wired Mag.* **14**(6) (2006). <http://www.wired.com/wired/archive/14.06/crowds.html>
30. Jarrar, M.: Position paper: towards the notion of gloss, and the adoption of linguistic resources in formal ontology engineering. In: *Proceedings of the 15th International Conference on World Wide Web. WWW '06*, pp. 497–503. ACM, New York (2006)

31. Jačala, M., Tvarožek, J.: Named entity disambiguation based on explicit semantics. In: Proceedings of the 38th International Conference on Current Trends in Theory and Practice of Computer Science, SOFSEM'12, pp. 456–466. Springer, Berlin, Heidelberg (2012)
32. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. *Knowl. Eng. Rev.* 18(1):1–31 (2003)
33. Köhler, J., Philippi, S., Specht, M., Rüegg, A.: Ontology based text indexing and querying for the semantic web. *Know. Based Syst.* 19(8), 744–754 (2006)
34. Kompan, M., Zeleník, D., Bielíková, M.: Methods for personalized recommendation of newspaper articles. In: *Znalosti (In Slovak)* (2011)
35. Kozareva, Z.: Bootstrapping named entity recognition with automatically generated gazetteer lists. In: Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research W. on - EACL '06, pp. 15–21. Association for Computational Linguistics, Morristown (2006)
36. Kramár, T., Barla, M., Bielíková, M.: Disambiguating search by leveraging the social network context based on the stream of user's activity. In: Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization, UMAP '10, pp. 387–392. Springer, Hawaii (2010)
37. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: Proceedings of Neural Information Processing Systems (NIPS). MIT Press, Cambridge (2003)
38. Lenat, D.B.: CYC: a large-scale investment in knowledge infrastructure. *Commun. ACM* 38(11), 33–38 (1995)
39. Liu, H., Singh, P.: Conceptnet—a practical commonsense reasoning tool-kit. *BT Technol. J.* 22(4), 211–226 (2004)
40. Liu, Q., Sung, A.H., Qiao, M.: Novel stream mining for audio steganalysis. In: Proceedings of the 17th ACM International Conference on Multimedia. MM '09, pp. 95–104. ACM, New York (2009)
41. Lu, L., Hanjalic, A.: Towards optimal audio “keywords” detection for audio content analysis and discovery. In: Proceedings of the 14th Annual ACM International Conference on Multimedia. MULTIMEDIA '06, pp. 825–834. ACM, New York (2006)
42. Magistrali, M., Catenazzi, N., Sommaruga, L.: Tonal mir: a music retrieval engine based on semantic web technologies. In: Proceedings of the 6th International Conference on Semantic Systems, I-SEMANTICS '10, pp. 21:1–21:5. ACM, New York (2010)
43. Maleewong, K., Anutariya, C., Wuwongse, V.: A semantic argumentation approach to collaborative ontology engineering. In: Proceedings of the 11th International Conference on Information Integration and Web-based Applications and Services. iiWAS '09, pp. 56–63. ACM, New York (2009)
44. Marchionini, G.: From finding to understanding. *Commun. ACM* 49(4), 41–46 (2006)
45. Mashhadi, A.J., Capra, L.: Quality control for real-time ubiquitous crowdsourcing. In: Proceedings of the 2nd International Workshop on Ubiquitous Crowdsourcing. UbiCrowd '11, pp. 5–8. ACM, New York (2011)
46. McDowell, L., Cafarella, M.: Ontology-driven, unsupervised instance population. *Web Semant. Sci. Serv. Agents World Wide Web* 6(3), 218–236 (2008)
47. Mizoguchi, R., Sunagawa, E., Kozaki, K., Kitamura, Y.: The model of roles within an ontology development tool: Hozo. *Appl. Ontol.* 2(2), 159–179 (2007)
48. Moor, A.D., Leenheer, P.D., Meersman, R., Starlab, V.: Dogma-mess: a meaning evolution support system for interorganizational ontology engineering. In: Proceedings of the 14th International Conference on Conceptual Structures, (ICCS 2006), pp. 189–203. Springer, Heidelberg (2006)
49. Mullins, M., Fizzano, P.: Treelicious: a system for semantically navigating tagged web pages. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 3, 91–96 (2010)
50. Orio, N.: Automatic identification of audio recordings based on statistical modeling. *Signal Process.* 90(4), 1064–1076 (2010)

51. Pantel, P., Pennacchiotti, M.: Automatically harvesting and ontologizing semantic relations. In: *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pp. 171–195. IOS Press, Amsterdam (2008)
52. Papadopoulos, G.T., Mylonas, P., Mezaris, V., Avrithis, Y.S., Kompatsiaris, I.: Knowledge-assisted image analysis based on context and spatial optimization. *Int. J. Semantic Web Inf. Syst.* **2**(3), 17–36 (2006)
53. Park, L.a.F., Ramamohanarao, K.: An analysis of latent semantic term self-correlation. *ACM Trans. Inf. Syst.* **27**(2), 1–35 (2009)
54. Parshotam, K.: Crowd computing: a literature review and definition. In: *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference. SAICSIT '13*, pp. 121–130. ACM, New York (2013)
55. Quinn, A.J., Bederson, B.B.: Human computation: a survey and taxonomy of a growing field. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '11*, pp. 1403–1412. ACM, New York (2011)
56. Radhakrishnan, R., Divakaran, A., Xiong, Z.: A time series clustering based framework for multimedia mining and summarization using audio features. In: *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval. MIR '04*, pp. 157–164. ACM, New York (2004)
57. Richter, S., Perkmann Berger, S., Koch, G., Füller, J.: Online idea contests: identifying factors for user retention. *Proceedings of the 5th International Conference on Online Communities and Social Computing. OCSC'13*, pp. 76–85. Springer, Berlin, Heidelberg (2013)
58. Sabou, M., Bontcheva, K., Scharl, A.: Crowdsourcing research opportunities: lessons from natural language processing. In: *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies, i-KNOW '12*, pp. 17:1–17:8. ACM, New York (2012)
59. Sanchez, D.: A methodology to learn ontological attributes from the web. *Data Knowl. Eng.* **69**(6), 573–597 (2010)
60. Sanchez, D., Moreno, A.: Learning non-taxonomic relationships from web documents for domain ontology construction. *Data Knowl. Eng.* **64**(3), 600–623 (2008)
61. Schedl, M., Widmer, G., Knees, P., Pohle, T.: A music information system automatically generated via web content mining techniques. *Inf. Process. Manage.* **47**(3), 426–439 (2011)
62. Siorpaes, K., Hepp, M.: Games with a purpose for the semantic web. *IEEE Intell. Syst.* **23**, 50–60 (2008)
63. Stewart, R., Scott, G., Zelevinsky, V.: Idea navigation: structured browsing for unstructured text. In: *Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, CHI '08*, pp. 1789–1792. ACM, New York (2008)
64. Tokarchuk, O., Cuel, R., Zamarian, M.: Analyzing crowd labor and designing incentives for humans in the loop. *IEEE Internet Comput.* **16**(5), 45–51 (2012)
65. Tsinaraki, C., Polydoros, P., Kazasis, F., Christodoulakis, S.: Ontology-based semantic indexing for mpeg-7 and tv-anytime audiovisual content. *Multimedia Tools Appl.* **26**(3), 299–325 (2005)
66. Tudorache, T., Noy, N.F., Falconer, S.M., Musen, M.A.: A knowledge base driven user interface for collaborative ontology development. *Proceedings of the 16th International Conference on Intelligent User Interfaces. IUI '11*, pp. 411–414. ACM, New York (2011)
67. Tvarožek, M.: Exploratory search in the adaptive social semantic web. *Inf. Sci. Technol. Bull. ACM Slovakia* **3**(1), 42–51 (2011)
68. Tvarožek, M., Bielíková, M.: Generating exploratory search interfaces for the semantic web. In: *Forbrig, P., Paternó, F., Mark Pejtersen, A. (eds.) Human-Computer Interaction, IFIP Advances in Information and Communication Technology*, vol. 332, pp. 175–186. Springer, Boston (2010)
69. Verborgh, R., Van Deursen, D., Mannens, E., Poppe, C., Van de Walle, R.: Enabling context-aware multimedia annotation by a novel generic semantic problem-solving platform. *Multimedia Tools Appl.* **61**(1), 105–129 (2012)
70. Wang, Y., Mei, T., Gong, S., Hua, X.S.: Combining global, regional and contextual features for automatic image annotation. *Pattern Recogn.* **42**(2), 259–266 (2009)

71. Weichselbraun, A., Wohlgenannt, G., Scharl, A.: Refining non-taxonomic relation labels with external structured data to support ontology learning. *Data Knowl. Eng.* **69**(8), 763–778 (2010)
72. Witbrock, M., Matuszek, C., Brusseau, A., Kahlert, R., Fraser, C.B., Lenat, D.: Knowledge begets knowledge: steps towards assisted knowledge acquisition in cyc. In: *Proceedings of the AAAI* (2005)
73. Zhu, S., Kane, S., Feng, J., Sears, A.: A crowdsourcing quality control model for tasks distributed in parallel. In: *CHI '12 Extended Abstracts on Human Factors in Computing Systems. CHI EA '12*, pp. 2501–2506. ACM, New York (2012)

Semantic Acquisition Games

Harnessing Manpower for Creating Semantics

Šimko, J.; Bieliková, M.

2014, IX, 141 p. 28 illus., Softcover

ISBN: 978-3-319-06114-6