Encyclopedic Knowledge Patterns from Wikipedia Links

Andrea Giovanni Nuzzolese[1,2], Aldo Gangemi[1], Valentina Presutti[1], and Paolo Ciancarini[1,2]

[1] STLab-ISTC Consiglio Nazionale delle Ricerche, Rome, Italy
[2] Dipartimento di Scienze dell'Informazione, Università di Bologna, Italy

Abstract. What is the most intuitive way of organizing concepts for describing things? What are the most relevant types of things that people use for describing other things? Wikipedia and Linked Data offer knowledge engineering researchers a chance to empirically identifying invariances in conceptual organization of knowledge i.e. knowledge patterns. In this paper, we present a resource of Encyclopedic Knowledge Patterns that have been discovered by analyzing the Wikipedia page links dataset, describe their evaluation with a user study, and discuss why it enables a number of research directions contributing to the realization of a meaningful Semantic Web.

1 Introduction

The realization of the Web of Data (aka Semantic Web) partly depends on the ability to make meaningful knowledge representation and reasoning. Elsewhere [5] we have introduced a vision of a pattern science for the Semantic Web as the means for achieving this goal. Such a science envisions the study of, and experimentation with, *knowledge patterns* (KP): small, well connected units of meaning which are 1) task-based, 2) well-grounded, and 3) cognitively sound. The first requirement comes from the ability of associating ontology vocabularies or schemas with explicit tasks, often called *competency questions* [7]: if a schema is able to answer a typical question an expert or user would like to make, it is a useful schema. The second requirement is related to the ability of ontologies to enable access to large data (which typically makes them successful) as well as being grounded in textual documents so as to support semantic technology applications that hybridize RDF data and textual documents. The third requirement comes from the expectation that schemas that more closely mirror the human ways of organizing knowledge are better. Unfortunately, evidence for this expectation is only episodic until now for RDF or OWL vocabularies [5].

Linked data and social web sources such as Wikipedia give us the chance to empirically study what are the patterns in organizing and representing knowledge i.e. knowledge patterns. KPs can be used for evaluating existing methods and models that were traditionally developed with a top-down approach, and open new research directions towards new reasoning procedures that better fit

L. Aroyo et al. (Eds.): ISWC 2011, Part I, LNCS 7031, pp. 520–536, 2011.

the actual Semantic Web applications need. In this study, we identify a set of invariances from a practical, crowd-sourced repository of knowledge: Wikipedia page links (wikilinks), which satisfy those three requirements, hence constituting good candidates as KPs. We call them Encyclopedic Knowledge Patterns (EKP) for emphasizing that they are grounded in encyclopedic knowledge expressed as linked data, i.e. DBpedia[1], and as natural language text, i.e. Wikipedia[2]. We have collected such set of EKPs in an open repository[3]. EKPs are able to answer the following (generic) competency question:

> *What are the most relevant entity types that provide an effective and intuitive description of entities of a certain type?*

For example, when describing "Italy" (a country), we typically indicate its neighbor countries, cities, administrative regions, spoken languages, etc. The EKP for describing countries should then include such a set of entity types: the most relevant for describing a country. We assume EKPs as cognitively sound because they emerge from the largest existing multi-domain knowledge source, collaboratively built by humans with an encyclopedic task in mind. This assumption is bound to our working hypothesis about the process of knowledge construction realized by the Wikipedia crowds: each article is linked to other articles when *explaining* or *describing* the entity referred to by the article. Therefore, the article's main subject can be said to be soundly and centrally related to the linked articles' main subjects. DBpedia, accordingly with this intuition, has rdf-ized a) the subjects referred to by articles as *resources*, b) the wikilinks as relations between those resources, and c) the types of the resources as OWL classes.

Hypotheses. Assuming that the articles linked from a Wikipedia page constitute a major source of descriptive knowledge for the subject of that page, we hypothesize that (i) the types of linked resources that occur most often for a certain type of resource constitute its EKP (i.e., the most relevant concepts to be used for describing resources of that type), and (ii) since we expect that any cognitive invariance in explaining/describing things is reflected in the wikilink graph, discovered EKPs are cognitively sound.

Contribution. The contribution of this paper is twofold: (i) we define an EKP discovery procedure, extract 184 EKPs, and publish them in OWL2 (ii) we support our hypotheses through a user-based evaluation, and discuss a number of research directions opened by our findings.

The paper is organized as follows: Section 2 discusses related work, Section 3 describes the resources we have used and the basic assumptions we have made, Section 4 focuses on the results we have gathered, Section 5 presents a user study for the evaluation and fine-tuning of EKPs, and Section 6 draws conclusions and gives an overview of research directions we are concentrating upon.

[1] http://dbpedia.org
[2] http://en.wikipedia.org
[3] The EKP repository is available at http://stlab.istc.cnr.it/stlab/WikiLinkPatterns

2 Related Work

To the best of our knowledge this work is the first attempt to extract knowledge patterns (KPs) from linked data. Nevertheless, there is valuable research on exploiting Wikipedia as a knowledge resource as well as on creating knowledge patterns.

Knowledge patterns. [5] argues that KPs are basic elements of the Semantic Web as an empirical science, which is the vision motivating our work. [4,16] present experimental studies on KPs, focusing on their creation and usage for supporting ontology design with shared good practices. Such KPs are usually stored in online repositories[4]. Contrary to what we present in this work, KPs are typically defined with a top-down approach, from practical experience in knowledge engineering projects, or extracted from existing, e.g. foundational, ontologies. These KPs are close to EKPs, but although some user-study proved that their use is beneficial in ontology design [3], yet they miss some of the aspects that we study here: evaluation of their cognitive soundness, and adequacy to provide access to large-scale linked data. [14] presents a resource of KPs derived from a lexical resource i.e., FrameNet [2]. In future work, we plan a compared analysis between EKPs and other KPs.

Building the web of data. Research focusing on feeding the Web of Data is typically centered on extracting knowledge from structured sources and transforming it into linked data. Notably, [8] describes how DBpedia is extracted from Wikipedia, and its linking to other Web datasets.

Another perspective is to apply knowledge engineering principles to linked data in order to improve its quality. [18] presents YAGO, an ontology extracted from Wikipedia categories and infoboxes that has been combined with taxonomic relations from WordNet. Here the approach can be described as a reengineering task for transforming a thesaurus, i.e. Wikipedia category taxonomy, to an ontology, which required accurate ontological analysis.

Extracting knowledge from wikipedia. Wikipedia is now largely used as a reference source of knowledge for empirical research. Research work from the NLP community, e.g., [20,9,15], exploits it as background knowledge for increasing the performance of algorithms addressing specific tasks. Two approaches are close to ours. [6] presents a method for inducing thesauri from Wikipedia by exploiting the structure of incoming wikilinks. The graph of wikilinks is used for identifying meaningful terms in the linked pages. In contrast, in our case we exploit outgoing wikilinks, as well as the full potential of the linked data semantic graph for identifying semantic entities as opposed to terms. [19] presents a statistical approach for the induction of expressive schemas for RDF data. Similarly to our study, the result is an OWL ontology, while in our experiment we extract novel schemas from wikilink structures that are encoded in RDF. Some studies have produced reusable results for improving the quality of the Web of

[4] E.g. the ontology design patterns semantic portal, http://www.ontologydesignpatterns.org

Data. We mention two notable examples: [13,1], which address the extraction of relations between Wikipedia entities, and [12] that presents a multi-lingual network of inter-connected concepts obtained by mining Wikipedia.

3 Materials and Methods

Our work grounds on the assumption that wikilink relations in DBpedia, i.e. instances of the `dbpo:wikiPageWikiLink` property[5], convey a rich encyclopedic knowledge that can be formalized as EKPs, which are good candidates as KPs [5].

Informally, an EKP is a small ontology that contains a concept S and its relations to the most relevant concepts C_j that can be used to describe S.

Representing invariances from wikipedia links. For representing wikilink invariances, we define *path* (type) as an extension of the notion of *property path*[6]:

Definition 1 (Path). *A path (type) is a property path (limited to length 1 in this work, i.e. a triple pattern), whose occurrences have (i) the same* `rdf:type` *for their subject nodes, and (ii) the same* `rdf:type` *for their object nodes. It is denoted here as:*

$$P_{i,j} = [S_i, p, O_j]$$

where S_i is a subject type, p is a property, and O_j is an object type of a triple. In this work, we only extract paths where p=`dbpo:wikiPageWikiLink`. Sometimes we use a simplified notation $[S_i, O_j]$, assuming $p =$ `dbpo:wikiPageWikiLink`.

We extract EKPs from paths (see Definition 2), however in order to formalize them, we perform a heuristic procedure to reduce multi-typing, to avoid redundancies, and to replace `dbpo:wikiPageWikiLink` with a contextualized object property. In practice, given a triple `s dbpo:wikiPageWikiLink o`, we construct its path as follows:

- the subject type S_i is set to the most specific type(s) of s
- the object type O_j is set to the most specific type(s) of o
- the property p of the path is set to the most general type of o

For example, the triple:

```
dbpedia:Andre_Agassi dbpo:wikiPageWikiLink dbpedia:Davis_Cup
```

would count as an occurrence of the following path:

$$Path_{Agassi,Davis} = [\texttt{dbpo:TennisPlayer}, \texttt{dbpo:Organisation}, \texttt{dbpo:TennisLeague}]$$

Figure 1 depicts such procedure for the path $Path_{Agassi,Davis}$:

[5] Prefixes `dbpo:`, `dbpedia:`, and `ka:` stand for `http:dbpedia.org/ontology/`, `http:dbpedia.org/resource/` and `http://www.ontologydesignpatterns.org/ont/lod-analysis-path.owl`, respectively.

[6] In SPARQL1.1 (http://www.w3.org/TR/sparql11-property-paths/) property paths can have length n, given by their route through the RDF graph.

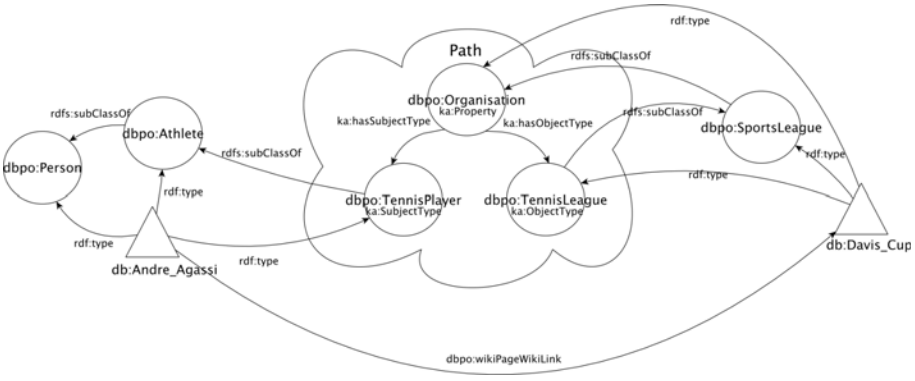

Fig. 1. Path discovered from the triple `dbpedia:Andre_Agassi dbpo:wikiPageWikiLink dbpedia:Davis_Cup`

- `dbpo:TennisPlayer` is the subject type because it is the most specific type of `dbpedia:Andre_Agassi`, i.e., `dbpo:TennisPlayer` $\sqsubseteq$ `dbpo:Person`;
- `dbpo:TennisLeague` is the object type because it is the most specific type of `dbpedia:Davis_Cup`, i.e., `dbpo:TennisLeague` $\sqsubseteq$ `dbpo:SportsLeague` $\sqsubseteq$ `dbpo:Organisation`
- `dbpo:Organisation` is the property of the path because it is the most general type of `dbpedia:Davis_Cup`.

Indicators. We use a set of indicators that are described in Table 1. Their application and related interpretation in this work are discussed in the following sections.

Table 1. Indicators used for empirical analysis of wikilink paths

Indicator	Description
$nRes(C)$	number of resources typed with a certain class C, $\lvert\{r_i$ `rdf:type` $C\}\rvert$
$nSubjectRes(P_{i,j})$	number of distinct resources that participate in a path as subjects, $\lvert\{(s_i$ `rdf:type` $S_i) \in P_{i,j} = [S_i, p, O_j]\}\rvert$
$pathPopularity(P_{i,j}, S_i)$	The ratio of how many distinct resources of a certain type participate as subject in a path to the total number of resources of that type. Intuitively, it indicates the popularity of a path for a certain subject type, $nSubjectRes(P_{i,j} = [S_i, p, O_j])$ divided by $nRes(S_i)$
$nPathOcc(P_{i,j})$	number of occurrences of a path $P_{i,j} = [S_i, p, O_j]$
$nPath(S_i)$	number of distinct paths having a same subject type S_i, e.g. the number of paths having `dbpo:TennisPlayer` as subject type
$AvPathOcc(S_i)$	sum of all $nPathOcc(P_{i,j})$ having a subject type S_i divided by $nPath(S_i)$ e.g. the avarage number of occurrences of paths having `dbpo:Philosopher` as subject type

Boundaries of Encyclopedic Knowledge Patterns. We choose the boundaries of an EKP by defining a threshold t for $pathPopularity(P_{i,j}, S_i)$. Accordingly, we give the following definition of $EKP(S_i)$ for a DBpedia type S_i.

Definition 2 (Encyclopedic Knowledge Patterns). *Let S_i be a DBpedia type, O_j $(j = 1, .., n)$ a list of DBpedia types, $P_{i,j} = [S_i, p, O_j]$ and t a threshold value.*
Given the triples:

```
dbpedia:s dbpedia-ont:wikiPediaWikiLink dbpedia:o
dbpedia:s rdf:type dbpedia:Si
dbpedia:o rdf:type dbpedia:Oj
```

we state that $EKP(S_i)$ is a set of paths, such that

$$P_{i,j} \in EKP(S_i) \iff pathPopularity(P_{i,j}, S_i) \geq t \quad (1)$$

We hypothesize values for t in Section 4, and evaluate them in Section 5.

OWL2 formalization of EKPs. We have stored paths and their associated indicators in a dataset, according to an OWL vocabulary called *knowledge architecture*[7]. Then, we have generated the Encyclopedic Knowledge Patterns (EKPs) repository[8] by performing a refactoring of the knowledge architecture data into OWL2 ontologies). Given the namespace `ekp:` and an $EKP(S_i) = [S_i, p_1, O_1], \ldots, [S_i, p_n, O_n]$, we formalize it in OWL2 by applying the following translation procedure:

- the name of the OWL file is `ekp:`[9] followed by the local name of S e.g., `ekp:TennisPlayer.owl`. Below we refer to the namespace of a specific EKP through the generic prefix `ekpS:`;
- S_i and O_j $j = 1, \ldots, n$ are refactored as `owl:Class` entities (they keep their original URI);
- p_j keep their original URI and are refactored as `owl:ObjectProperty` entities;
- for each O_j we create a sub-property of p_{i+n}, $ekpS : O_j$ that has the same local name as O_j and the `ekpS:` namespace; e.g. `ekp:TennisPlayer.owl#TennisLeague`.
- for each $ekpS : O_j$ we add an `owl:allVauesFrom` restriction to S_i on $ekpS : O_j$, with range O_j.

For example, if $Path_{Agassi,Davis}$ (cf. Figure 1) is part of an EKP, it gets formalized as follows:

```
Prefix: dbpo: <http://dbpedia.org/ontology/>
Prefix:
   ekptp: <http://www.ontologydesignpatterns.org/ekp/TennisPlayer.owl#>
Ontology: <http://www.ontologydesignpatterns.org/ekp/TennisPlayer.owl>
Class: dbpo:TennisPlayer
   SubClassOf:
```

[7] http://www.ontologydesignpatterns.org/ont/lod-analysis-path.owl
[8] The EKP repository is available at http://stlab.istc.cnr.it/stlab/WikiLinkPatterns.
[9] The prefix `ekp:` stands for the namespace http://www.ontologydesignpatterns.org/ekp/.

Table 2. Dataset used and associated figures

Dataset	Description	Indicator	Value
DBPO	DBpedia ontology	Number of classes	272
dbpedia_instance_types_en	Resource types i.e. rdf:type triples	Number of resources having a DBPO type rdf:type triples	1,668,503 6,173,940
dbpedia_page_links_en	Wikilinks triples	Number of resources used in wikilinks Number of wikilinks	15,944,381 107,892,317
DBPOwikilinks	Wikilinks involving only resources typed with DBPO classes	Number of resources used in wikilinks Number of wikilinks	1,668,503 16,745,830

```
      ekptp:TennisLeague only dbpo:TennisLeague
Class: dbpo:TennisLeague
ObjectProperty: ekptp:TennisLeague
   SubPropertyOf: dbpo:Organisation
...
```

Materials. We have extracted EKPs from a subset of the DBpedia wikilink dataset (*dbpedia_page_links_en*), and have created a new dataset (*DBPOwikilinks*) including only links between resources that are typed by DBpedia ontology version 3.6 (DBPO) classes (15.52% of the total wikilinks in *dbpedia_page_links_en*). *DBPOwikilinks* excludes a lot of links that would create semantic interpretation issues, e.g. images (e.g. `dbpedia:Image:Twitter_2010_logo.svg`), Wikipedia categories (e.g. `dbpedia:CAT:Vampires_in_comics`), untyped resources (e.g. `dbpedia:%23Drogo`), etc.

DBPO currently includes 272 classes, which are used to type 10.46% of the resources involved in *dbpedia_page_links_en*. We also use *dbpedia_instance_types_en*, which contains type axioms, i.e. `rdf:type` triples. This dataset contains the materialization of all inherited types (cf. Section 4). Table 2 summarizes the figures described above.

4 Results

We have extracted 33,052 paths from the English wikilink datasets, however many of them are not relevant either because they have a limited number of occurrences, or because their subject type is rarely used. In order to select the paths useful for EKP discovery (our goal) we have considered the following criteria:

– *Usage in the wikilink dataset.* The resources involved in *dbpedia_page_links_en* are typed with any of 250 DBPO classes (out of 272). Though, we are interested in *direct types*[10] of resources in order to avoid redundancies when counting path occurrences. For example, the resource `dbpedia:Ludwik_Fleck` has three types `dbpo:Scientist;dbpo:Person;owl:Thing` because type assertions in DBpedia are materialized along the hirerachy of DBPO. Hence, only

[10] In current work, we are also investigating indirectly typed resource count, which might lead to different EKPs, and to empirically studying EKP ordering.

`dbpo:Scientist` is relevant to our study. Based on this criterion, we keep only 228 DBPO classes and the number of paths decreases to 25,407.

- *Number of resources typed by a class C (i.e., nRes(C)).* Looking at the distribution of resource types, we have noticed that 99.98% of DBPO classes have at least 30 resource instances. Therefore we have decided to keep paths whose subject type C has at least $nRes(C)=30$.
- *Number of path occurrences having a same subject type (i.e., $nPathOcc(P_{i,j})$).* The average number of outgoing wikilinks per resource in *dbpedia_page_links_en* is 10. Based on this observation and on the previous criterion, we have decided to keep paths having at least $nPathOcc(P)=30*10=300$.

After applying these two criteria, only 184 classes and 21,503 paths are retained. For example, the path [Album,Drug] has 226 occurrences, and the type `dbpo:AustralianFootballLeague` has 3 instances, hence they have been discarded.

EKP discovery. At this point, we had each of the 184 classes used as subject types associated with a set of paths, each set with a cardinality ranging between 2 and 191 (with 86.29% of subjects bearing at least 20 paths). Our definition of EKP requires that its backbone be constituted of a small number of object types, typically below 10, considering the existing resources of models that can be considered as KPs (see later in this section for details). In order to generate EKPs from the extracted paths, we need to decide what threshold should be used for selecting them, which eventually creates appropriate *boundaries* for EKPs. In order to establish some meaningful threshold, we have computed the ranked distributions of $pathPopularity(P_{i,j}, S_i)$ for each selected subject type, and measured the correlations between them. Then, we have fine-tuned these findings by means of a user study (cf. Section 5), which had the dual function of both evaluating our results, and suggesting relevance criteria for generating the EKP resource. Our aim is to build a *prototypical* ranking of the $pathPopularity(P_{i,j}, S_i)$ of the selected 184 subject types, called $pathPopularity_{DBpedia}$, which should show how relevant paths for subject types are typically distributed according to the Wikipedia crowds, hence allowing us to propose a threshold criterion for any subject type. We have proceeded as follows.

1. We have chosen the top-ranked 40 paths ($P_{i,j}$) for each subject type (S_i), each constituting a $pathPopularity(P_{i,j}, S_i)$. Some subject types have less than 40 paths: in such cases, we have added 0 values until filling the gap. The number 40 has been chosen so that it is large enough to include not only paths covering at least 1% of the resources, but also much rarer ones, belonging to the long tail.
2. In order to assess if a prototypical ranking $pathPopularity_{DBpedia}$ would make sense, we have performed a multiple correlation between the different $pathPopularity(P_{i,j}, S_i)$. In case of low correlation, the prototypical ranking would create odd effects when applied to heterogeneous rank distributions across different S_i. In case of high correlation, the prototype would make

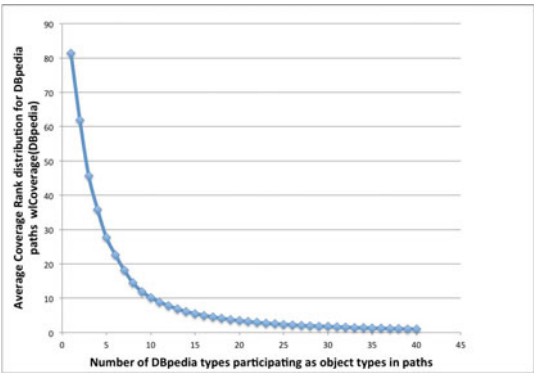

Fig. 2. Distribution of $pathPopularity_{DBpedia}$: the average values of popularity rank i.e., $pathPopularity(P_{i,j}, S_i)$, for DBpedia paths. The x-axis indicates how many paths (on average) are above a certain value t of $pathPopularity(P, S)$.

sense, and we can get reassured that the taxonomy we have used (DBPO in this experiment) nicely fits the way wikilinks are created by the Wikipedia crowds.

3. We have created a prototypical distribution $pathPopularity_{DBpedia}$ that is representative for all S_i distributions. Such a distribution is then used to hypothesize some thresholds for the relevance of $P_{i,j}$ when creating boundaries for EKPs. The thresholds are used in Section 5 to evaluate the proposed EKPs with respect to the rankings produced during the user study.

In order to measure the distribution from step 2, we have used the Pearson correlation measure ρ, ranging from -1 (no agreement) to +1 (complete agreement), between two variables X and Y i.e. for two different S_i in our case. The correlation has been generalized to all 16,836 pairs of the 184 $pathPopularity(P_{i,j}, S_i)$ ranking sets ($184 * 183/2$), in order to gather a multiple correlation. The value of such multiple correlation is 0.906, hence excellent.

Once reassured on the stability of $pathPopularity(P_{i,j}, S_i)$ across the different S_i, we have derived (step 3) $pathPopularity_{DBpedia}$, depicted in Figure 2.

In order to establish some reasonable relevance thresholds, $pathPopularity_{DBpedia}$ has been submitted to K-Means Clustering, which generates 3 small clusters with popularity ranks above 22.67%, and 1 large cluster (85% of the 40 ranks) with popularity ranks below 18.18%. The three small clusters includes seven paths: this feature supports the buzz in cognitive science about a supposed amount of 7 ± 2 objects that are typically manipulated by the cognitive systems of humans in their recognition tasks [11,10]. While the 7 ± 2 conjecture is highly debated, and possibly too generic to be defended, this observation has been used to hypothesize a first threshold criterion: since the seventh rank is at 18.18% in $pathPopularity_{DBpedia}$, this value of $pathPopularity(P_{i,j}, S_i)$ will be our first guess for including a path in an EKP. We propose a second threshold based on FrameNet [2], a lexical database, grounded in a textual corpus, of situation types called *frames*. FrameNet is

Table 3. Sample paths for the subject type `Album`: number of path occurrences, distinct subject resources, and popularity percentage value

Path	$nPathOcc(P_{i,j})$	$nSubjectRes(P_{i,j})$	$pathPopularity(P_{i,j}, S_i)$ (%)
`[Album,Album]`	170,227	78,137	78.89
`[Album,MusicGenre]`	108,928	68,944	69.61
`[Album,MusicalArtist]`	308,619	68,930	69.59
`[Album,Band]`	125,919	62,762	63.37
`[Album,Website]`	62,772	49,264	49.74
`[Album,RecordLabel]`	56,285	47,058	47.51
`[Album,Single]`	114,181	29,051	29.33
`[Album,Country]`	40,296	25,430	25.67

currently the only cognitively-based resource of potential knowledge patterns (the frames, cf. [14]). The second threshold (11%) is provided by the average number of frame elements in FrameNet frames (frame elements roughly correspond to paths for EKPs), which is 9 (the ninth rank in $pathPopularity_{DBpedia}$ is at 11%). The mode value of frame elements associated with a frame is 7, which further supports our proposal for the first threshold. An example of the paths selected for a subject type according to the first threshold is depicted in Tab. 3, where some paths for the type `Album` are ranked according to their $pathPopularity(P_{i,j}, S_i)$. In Section 5 we describe an evaluation of these threshold criteria by means of a user study.

Threshold criteria are also used to enrich the formal interpretation of EKPs. Our proposal, implemented in the OWL2 EKP repository, considers the first threshold as an indicator for an existential quantification over an OWL restriction representing a certain path. For example, `[Album,MusicGenre]` is a highly-popular path in the `Album` EKP. We interpret high-popularity as a feature for generating an existential interpretation, i.e.: `Album` $\sqsubseteq$ ($\exists$`MusicGenre.MusicGenre`). This interpretation suggests that each resource typed as an `Album` has at least one `MusicGenre`, which is intuitively correct. Notice that even if all paths have a $pathPopularity(P_{i,j}, S_i)$ of less that 100%, we should keep in mind that semantic interpretation over the Web is made in open-world, therefore we feel free to assume that such incompleteness is a necessary feature of Web-based knowledge (and possibly of any crowd-sourced knowledge).

5 Evaluation

Although our empirical observations on DBpedia could give us means for defining a value for the threshold t (see Definition 2 and Section 4), we still have to prove that emerging EKPs provide an intuitive schema for organizing knowledge. Therefore, we have conducted a user study for making users identify the EKPs associated with a sample set of DBPO classes, and for comparing them with those emerging from our empirical observations.

User study. We have selected a sample of 12 DBPO classes that span social, media, commercial, science, technology, geographical, and governmental

Table 4. DBPO classes used in the user-study and their related figures

DBPO class type	**nRes(S)**	$nPath(S_i)$	$AvPathOcc(S_i)$
Language	3,246	99	29.27
Philosopher	1,009	112	18.29
Writer	10,102	172	15.30
Ambassador	286	85	15.58
Legislature	453	83	25.11
Album	99,047	172	11.71
Radio Station	16,310	151	7.31
Administrative Region	31,386	185	11.30
Country	2,234	169	35.16
Insect	37,742	98	9.16
Disease	5,215	153	12.10
Aircraft	6,420	126	10.32

domains. They are listed in Table 4. For each class, we indicate the number of its resources, the number of paths it participates in as subject type, and the average number of occurrences of its associated paths. We have asked the users to express their judgement on how relevant were a number of (object) types (i.e., paths) for describing things of a certain (subject) type. The following sentence has been used for describing the user study task to the users:

> *We want to study the best way to describe things by linking them to other things. For example, if you want to describe a person, you might want to link it to other persons, organizations, places, etc. In other words, what are the most relevant types of things that can be used to describe a certain type of things?*

We asked the users to fill a number of tables, each addressing a class in the sample described in Table 4. Each table has three columns:

- *Type 1* indicating the class of things (subjects) to be described e.g. `Country`;
- A second column to be filled with a relevance value for each row based on a scale of five relevance values, Table 5 shows the scale of relevance values and their interpretations as they have been provided to the users. Relevance values had to be associated with each element of Type 2;
- *Type 2* indicating a list of classes of the paths (i.e. the object types) in which *Type 1* participates as subject type. These were the suggested types of things that can be linked for describing entities of *Type 1* e.g. `Administrative Region`, `Airport`, `Book`, etc.

By observing the figures of DBPO classes (cf. Table 4) we realized that the entire list of paths associated with a subject type would have been too long to be proposed to the users. For example, if *Type 1* was `Country`, the users would have been submitted 169 rows for *Type 2*. Hence, we decided a criterion for selecting a representative set of such paths. We have set a value for t to 18% and have included, in the sample set, all $P_{i,j}$ such that $pathPopularity(P_{i,j}, S_i) \geq 18\%$. Furthermore, we have also included an additional random set of 14 $P_{i,j}$ such that $pathPopularity(P_{i,j}, S_i) < 18\%$.

We have divided the sample set of classes into two groups of 6. We had ten users evaluating one group, and seven users evaluating the other group. Notice

Table 5. Ordinal (Likert) scale of relevance scores

Relevance score	Interpretation
1	The type is irrelevant;
2	The type is slightly irrelevant;
3	I am undecided between 2 and 4;
4	The type is relevant but can be optional;
5	The type is relevant and should be used for the description.

Table 6. Average coefficient of concordance for ranks (Kendall's W) for the two groups of users

User group	Average inter-rater agreement
Group 1	0.700
Group 2	0.665

that the users come from different cultures (Italy, Germany, France, Japan, Serbia, Sweden, Tunisia, and Netherlands), and speak different mother tongues. In practice, we wanted to avoid focusing on one specific language or culture, at the risk of reducing consensus.

In order to use the EKPs resulting from the user study as a reference for next steps in our evaluation task, we needed to check the inter-rater agreement. We have computed the Kendall's coefficient of concordance for ranks (W), for all analyzed DBPO classes, which calculates agreements between 3 or more rankers as they rank a number of subjects according to a particular characteristic. Kendall's W ranges from 0 (no agreement) to 1 (complete agreement). Table 6 reports such values for the two groups of users, which show that we have reached a good consensus in both cases. Additionally, Table 7 reports W values for each class in the evaluation sample.

Table 7. Inter-rater agreement computed with Kendall's W (for all values $p < 0.0001$) and reliability test computed with Cronbach's alpha

DBPO class	Agreement	Reliability	DBPO class	Agreement	Reliability
Language	0.836	0.976	Philosopher	0.551	0.865
Writer	0.749	0.958	Ambassador	0.543	0.915
Legislature	0.612	0.888	Album	0.800	0.969
Radio Station	0.680	0.912	Administrative Region	0.692	0.946
Country	0.645	0.896	Insect	0.583	0.929
Disease	0.823	0.957	Aircraft	0.677	0.931

Evaluation of emerging DBpedia EKPs through correlation with user-study results: how good is DBpedia as a source of EKPs? The second step towards deciding t for the generation of EKPs has been to compare DBpedia EKPs to those emerging from the users' choices. DBpedia $EKP(S_i)$ would result from a selection of paths having S_i as subject type, based on their associated $pathPopularity(P_{i,j}, S_i)$ values (to be $\geq t$). We had to compare the $pathPopularity(P_{i,j}, S_i)$ of the paths associated with the DBPO sample classes (cf. Table 4), to the relevance scores assigned by the users. Therefore, we needed to define a mapping function between $pathPopularity(P_{i,j}, S_i)$ values and the 5-level scale of relevance scores (Table 5).

We have defined the mapping by splitting the $pathPopularity_{DBpedia}$ distribution (cf. Figure 2) into 5 intervals, each corresponding to the 5 relevance scores of the Likert scale used in the user-study. Table 8 shows our hypothesis

Table 8. Mapping between $wlCoverage_{DBpedia}$ intervals and the relevance score scale

$pathPopularity_{DBpedia}$ **interval**	**Relevance score**
[18, 100]	5
[11, 18[	4
]2, 11[	3
]1, 2]	2
[0, 1]	1

Table 9. Average multiple correlation (Spearman ρ) between users' assigned scores, and $pathPopularity_{DBpedia}$ based scores

User group	**Correl. with DBpedia**
Group 1	0.777
Group 2	0.717

Table 10. Multiple correlation coefficient (ρ) between users's assigned score, and $pathPopularity_{DBpedia}$ based score

DBPO class	**Correl. users / DBpedia**	**DBPO class**	**Correl. users / DBpedia**
Language	0.893	Philosopher	0.661
Writer	0.748	Ambassador	0.655
Legislature	0.716	Album	0.871
Radio Station	0.772	Administrative Region	0.874
Country	0.665	Insect	0.624
Disease	0.824	Aircraft	0.664

of such mapping. The hypothesis is based on the thresholds defined in Section 4. The mapping function serves our purpose of performing the comparison and identifying the best value for t, which is our ultimate goal. In case of scarce correlation, we expected to fine-tune the intervals for finding a better correlation and identifying the best t. Based on the mapping function, we have computed the relevance scores that DBpedia would assign to the 12 sample types, and calculated the Spearman correlation value (ρ) wich ranges from -1 (no agreement) to $+1$ (complete agreement) by using the *means* of relevance scores assigned by the users. This measure gives us an indication on how precisely DBpedia wikilinks allow us to identify EKPs as compared to those drawn by the users. As shown in Table 9, there is a good correlation between the two distributions. Analogously, Table 10 shows the multiple correlation values computed for each class, which are significantly high. Hence, they indicate a satisfactory precision.

We can conclude that our hypothesis (cf. Section 1) is supported by these findings, and that Wikipedia wikilinks are a good source for EKPs. We have tested alternative values for t, and we have found that our hypothesized mapping (cf. Table 8) provides the best correlation values among them. Consequently, we have set the threshold value for EKP boundaries (cf. Definition 2) as $t = 11\%$.

6 Discussion and Conclusions

We have presented a study for discovering Encyclopedic Knowledge Patterns (EKP) from Wikipedia page links. In this study we have used the DBPO classes to create a wikilink-based partition of crowd-sourced encyclopedic knowledge expressed as paths of length 1, and applied several measures to create a *boundary*

around the most relevant object types for a same subject type out of wikilink triples.

Data have been processed and evaluated by means of both statistical analysis over the paths, and a user study that created a reference ranking for a subset of subject types and their associated paths. Results are very good: *stable criteria for boundary creation* (high correlation of path popularity distributions across subject types), *large consensus* among (multicultural) users, and *good precision* (high correlation between users' and EKP rankings).

The 184 EKP so generated have been formalized in OWL2 and published, and can be used either as lenses for the exploration of DBpedia, or for designing new ontologies that inherit the data and textual grounding provided by DBpedia and Wikipedia. Also data linking can take advantage of EKPs, by modularizing the datasets to be linked.

There are many directions that the kind of research we have done opens up: some are presented in the rest of this section.

Applying EKPs to resource concept maps. An application of EKPs is the creation of synthetic concept maps out of the wikilinks of a resource. For example, a concept map of all wikilinks for the resource about the scientist `dbpr:Ludwik_Fleck` contains 44 unordered resources, while a concept map created with a lens provided by the `Scientist` EKP provides the 13 most typical resources with explicit relations. We should remark that EKPs typically (and intentionally) exclude the "long tail" features of a resource, which sometimes are important. Investigating how to make these relevant long tail features emerge for specific resources and requirements is one of the research directions we want to explore.

Wikilink relation semantics. An obvious elaboration of EKP discovery is to infer the object properties that are implicit in a wikilink. This task is called *relation discovery*. Several approaches have been used for discovering relations in Wikipedia, (cf. Section 2, [9] is an extensive overview), and are being investigated. Other approaches are based on the existing semantic knowledge from DBpedia: three of them are exemplified here because their results have already been implemented in the EKP resource.

Induction from infobox properties. For example, the path `[Album,MusicalArtist]` features a distribution of properties partly reported in Table 11. There is a clear majority for the `producer` property, but other properties are also present, and some are even clear anomalies (e.g. `*[Album,dbprop:nextAlbum,MusicalArtist]`[11]). In general, there are two typical situations: the first is exemplified by `[Album,MusicalArtist]`, where the most frequent property covers only part of the possible semantics of the wikilink paths. The second situation is when the most frequent property is maximally general, and repeats the name of the object type, e.g. `[Actor,dbprop:film,Film]`. In our EKP resource, we add the most frequent

[11] * indicates a probably wrong path.

Table 11. Sample paths for the subject type `Album` from the infobox DBpedia dataset, with their frequency. Some paths are clear mistakes.

Path	$nPathOcc(P_{i,j})$
`[Album,dbprop:producer,MusicalArtist]`	3,413
`[Album,dbprop:artist,MusicalArtist]`	236
`[Album,dbprop:writer,MusicalArtist]`	46
`[Album,dbprop:lastAlbum,MusicalArtist]`	35
`*[Album,dbprop:nextAlbum,MusicalArtist]`	33
`[Album,dbprop:thisAlbum,MusicalArtist]`	27
`[Album,dbprop:starring,MusicalArtist]`	20

properties from the infobox dataset as annotations, accompanied by a frequency attribute.

Induction from top superclasses. For example, the path `[Album,MusicalArtist]` can be enhanced by inducing the top superclass of `MusicalArtist`, i.e. `Person`, as its property. This is possible either in RDFS, or in OWL2 (via punning). The path would be in this case `[Album,Person,MusicalArtist]`. This solution has not precision problems, but is also quite generic on the semantics of a wikilink.

Punning of the object type. For example, the path `[Album,MusicalArtist]` can be enriched as `[Album,MusicalArtist,MusicalArtist]`. This solution is pretty uninformative at the schema level, but can be handy when an EKP is used to visualize knowledge from wikilinks, for example in the application described above of a resource concept map, where resources would be linked with the name of the object type: this results to be very informative for a concept map user. In our EKP resource, we always reuse the object type as a (locally defined) object property as well.

Additional approaches we have conceived would exploit existing resources created by means of NLP techniques (e.g. WikiNet, [12]), or by game-based crowdsourcing (e.g. OpenMind [17]).

Intercultural issues. Given the multilingual and multicultural nature of Wikipedia, comparison between EKPs extracted from different versions of Wikipedia is very interesting. We have extracted EKPs from English and Italian versions, and we have measured the correlation between some English- and Italian-based EKPs. The results are encouraging; e.g. for the subject type `Album` the Spearman correlation between the top 20 paths for Italian resources and those for English ones is 0.882%, while for `Language` is 0.657%. This despite the fact that Italian paths have lower popularity values than English ones, and much fewer wikilinks (3.25 wikilinks per resource on average).

Schema issues. DBPO has been generated from Wikipedia infoboxes. The DBpedia infobox dataset contains 1,177,925 object property assertions, and their objects are also wikilinks. This means that 7.01% of the 16,745,830 wikilink triples that we have considered overlap with infobox triples. This is a potential bias on our results, which are based on DBPO; however, such bias is very limited:

removing 7% of the wikilinks is not enough to significantly decrease the high correlations we have found.

Finally, one might wonder if our good results could be obtained by using other ontologies instead of DBPO. We are experimenting with wikilink paths typed by Yago [18], which has more than 288,000 classes, and a broader coverage of resources (82% vs. 51.9% of DBPO). Working with Yago is very interesting, but also more difficult, since it applies multityping extensively, and the combinatorics of its paths is orders of magnitude more complex than with paths typed by DBPO. Sample Yago paths include e.g.: `[Coca-ColaBrands,BoycottsOfOrganizations]`, `[DietSodas,LivingPeople]`. Those paths are domain-oriented, which is a good thing, but they also share a low popularity (ranging around 3% in top ranks) in comparison to DBPO classes. In other words, the skewness of Yago $pathPopularity(P_{i,j}, S_i)$ is much higher than that of DBPO, with a very long tail. However, the clustering factor is not so different: a Yago EKP can be created e.g. for the class `yago:AmericanBeerBrands`, and its possible thresholds provided by K-Means Clustering appear very similar to the ones found for DBPO EKPs: we should only scale down the thresholds, e.g. from 18% to 1%.

Acknowledgements. This work has been part-funded by the European Commission under grant agreement FP7-ICT-2007-3/ No. 231527 (IKS - Interactive Knowledge Stack). We would like to thank Milan Stankovic for his precious advise.

References

1. Akbik, A., Broß, J.: Wanderlust: Extracting Semantic Relations from Natural Language Text Using Dependency Grammar Patterns. In: Proc. of the Workshop on Semantic Search (SemSearch 2009) at the 18th International World Wide Web Conference (WWW 2009), Madrid, Spain, pp. 6–15 (2009)
2. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet Project. In: Proc. of the 17th International Conference on Computational Linguistics, Morristown, NJ, USA, pp. 86–90 (1998)
3. Blomqvist, E., Presutti, V., Gangemi, A.: Experiments on pattern-based ontology design. In: Proceeding of K-CAP 2009, Redondo Beach, California, USA, pp. 41–48. ACM (2009)
4. Blomqvist, E., Sandkuhl, K., Scharffe, F., Svatek, V.: Proc. of the Workshop on Ontology Patterns (WOP, collocated with the 8th International Semantic Web Conference (ISWC-2009), Washington D.C., USA. CEUR Workshop Proceedings, vol. 516 (October 25 (2009)
5. Gangemi, A., Presutti, V.: Towards a Pattern Science for the Semantic Web. Semantic Web 1(1-2), 61–68 (2010)
6. Giuliano, C., Gliozzo, A.M., Gangemi, A., Tymoshenko, K.: Acquiring Thesauri from Wikis by Exploiting Domain Models and Lexical Substitution. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010. LNCS, vol. 6089, pp. 121–135. Springer, Heidelberg (2010)

7. Gruninger, M., Fox, M.S.: The role of competency questions in enterprise engineering. In: Proc. of the IFIP WG5.7 Workshop on Benchmarking - Theory and Practice, Trondheim, Norway, pp. 83–95 (1994)
8. Lehmann, J., Bizer, C., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A Crystallization Point for the Web of Data. Journal of Web Semantics 7(3), 154–165 (2009)
9. Medelyan, O., Milne, D., Legg, C., Witten, I.H.: Mining meaning from wikipedia. Int. J. Hum.-Comput. Stud. 67(9), 716–754 (2009)
10. Migliore, M., Novara, G., Tegolo, D.: Single neuron binding properties and the magical number 7. Hippocampus 18(11), 1122–1130 (2008)
11. Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review 63(2), 81–97 (1956)
12. Nastase, V., Strube, M., Boerschinger, B., Zirn, C., Elghafari, A.: WikiNet: A Very Large Scale Multi-Lingual Concept Network. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proc. of the Seventh International Conference on Language Resources and Evaluation (LREC), pp. 1015–1022. European Language Resources Association (2010)
13. Nguyen, D.P.T., Matsuo, Y., Ishizuka, M.: Relation extraction from wikipedia using subtree mining. In: Proc. of the 22nd National Conference on Artificial Intelligence, vol. 2, pp. 1414–1420. AAAI Press (2007)
14. Nuzzolese, A.G., Gangemi, A., Presutti, V.: Gathering Lexical Linked Data and Knowledge Patterns from FrameNet. In: Proc. of the 6th International Conference on Knowledge Capture (K-CAP), Banff, Alberta, Canada, pp. 41–48 (2011)
15. Ponzetto, S.P., Navigli, R.: Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia.. In: Boutilier, C. (ed.) IJCAI, Pasadena, USA, pp. 2083–2088 (2009)
16. Presutti, V., Chaudhri, V.K., Blomqvist, E., Corcho, O., Sandkuhl, K.: Proc. of the Workshop on Ontology Patterns (WOP 2010) at ISWC-2010, Shangai, China. CEUR Workshop Proceedings (November 8, 2010)
17. Singh, P.: The Open Mind Common Sense project. Technical report, MIT Media Lab (2002)
18. Suchanek, F., Kasneci, G., Weikum, G.: Yago - A Large Ontology from Wikipedia and WordNet. Elsevier Journal of Web Semantics 6(3), 203–217 (2008)
19. Völker, J., Niepert, M.: Statistical Schema Induction. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011, Part I. LNCS, vol. 6643, pp. 124–138. Springer, Heidelberg (2011)
20. Zesch, T., Müller, C., Gurevych, I.: Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary.. In: Proc. of the Sixth International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco, pp. 1646–1652. European Language Resources Association (2008)