WiGipedia: A Tool for Improving Structured Data in Wikipedia

Svetlin Bostandjiev[†], John O'Donovan, Christopher Hall, Brynjar Gretarsson, Tobias Höllerer[‡]
Computer Science Department
University of California Santa Barbara
[†]alex@cs.ucsb.edu, [‡]holl@cs.ucsb.edu

***Abstract*—Wikipedia is emerging as the dominant global knowledge repository. Recently, large numbers of users have collaborated to produce more structured information in the so called "infoboxes". However, editing this data requires even more care than editing standard wikitext, as one must follow arcane template syntax. This paper describes *WiGipedia*, a novel tool which provides an alternative to the traditional approach, by supporting editing of structured wiki data through two intuitive and interactive interfaces, facilitating user input on both tabular and graph-based representations of structured data. The tool allows users to identify and correct inconsistencies that are otherwise hidden across multiple articles. Furthermore, a novel recommendation algorithm is applied to assist users in their contribution to the wiki. The paper discusses design, implementation details, and results of a usability study in which the system compares significantly well against the traditional approach to editing Wikipedia infoboxes.**

I. Introduction

Today, Wikipedia is one of the largest community created data sets on the social web. However, over the last five years there has been a significant falloff in the number of users providing meaningful contributions [1]. While some of this falloff can be attributed to more complex security policies and saturation of simpler articles, it remains that the current MediaWiki interface and markup is complicated for an average user to provide contributions. This paper focuses on two core contributions, operating in the process flow of Figure 1. Firstly, provision of an easy-to-use tool to elicit content from viewers for specific Wikipedia structured information. Secondly, using that content to boost the *consistency* of structured data spread across multiple Wikipedia articles.

We believe that while automated inference algorithms such as Kylin [2] and Kog [3] can help to fill in the enormous gaps in the semantic graph of Wikipedia through techniques such as subsumption detection, it is important to keep the user "in-the-loop" to encourage addition of new raw data and maintain healthy growth. We present *WiGipedia*[1], an interactive interface which can be embedded in a wiki article web page. The interface can represent related data as either a graph or a table. Data is sourced from within the target article and a subset of its semantically connected articles via queries to DBpedia [4], as illustrated in Figure 2.

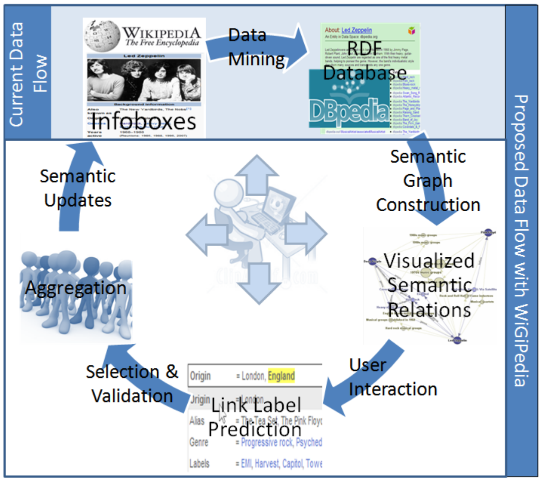


Fig. 1. Process flow in *WiGipedia*. The shaded area shows existing information flow between Wikipedia and DBpedia, while the lighter area shows the new steps facilitated by our system. At each step, *WiGipedia* keeps real users involved in data analysis.

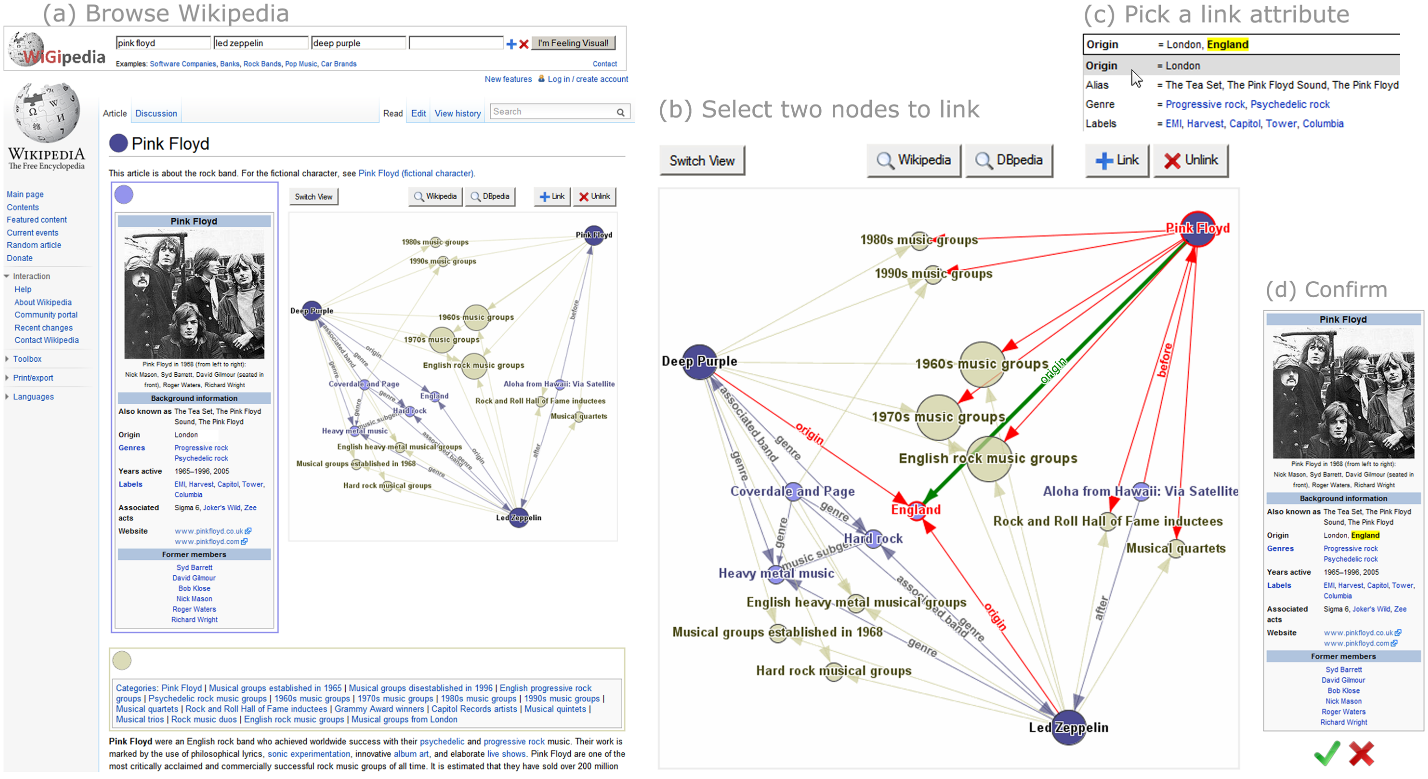
A. Inconsistent Data

DBpedia is a semantic web resource which crawls structured data from Wikipedia and organizes it into a database of subject-object-predicate triples. The resulting semantic data suffers from incompleteness and a variety of inconsistencies, which adversely impact results of complex queries over the data. This inconsistency is partly due to the manner in which users edit Wikipedia. Since users can only view one article at a time, it becomes difficult for them to recognize what information is missing or incorrect relative to other wiki articles. We believe that by allowing users to edit related information *spread across multiple articles* in a single view, we are taking initial steps towards correcting these global inconsistencies in Wikipedia.

For the purpose of our discussions, we can loosely classify data inconsistencies in Wikipedia as follows:

1) *Ambiguity*: Standard double meanings of terms. E.g.: A record company has a "Band" (Music), and an atom has a "Band" (Electrons).
2) *Multiple Naming*: Entities can have multiple names or aliases. E.g.: United Kingdom, U.K., UK.
3) *Misspelling*: Spelling errors exist on some entities.
4) *Inconsistent links*: When multiple edges incumbent on a

[1] http://www.wigipedia-online.com



Examples: Software Companies, Banks, Rock Bands, Pop Music, Car Brands

Pink Floyd

This article is about the rock band. For the fictional character, see Pink Floyd (fictional character).

Pink Floyd

Pink Floyd in 1968 (from left to right): Nick Mason, Syd Barrett, David Gilmour (seated in front), Roger Waters, Richard Wright

Background information

Also known as	The Tea Set, The Pink Floyd Sound, The Pink Floyd
Origin	London
Genres	Progressive rock Psychedelic rock
Years active	1965–1996, 2005
Labels	EMI, Harvest, Capitol, Tower, Columbia
Associated acts	Sigma 6, Joker's Wild, Zee
Website	www.pinkfloyd.co.uk www.pinkfloyd.com

Former members

Syd Barrett
David Gilmour
Bob Klose
Nick Mason
Roger Waters
Richard Wright

Categories: Pink Floyd | Musical groups established in 1965 | Musical groups disestablished in 1996 | English progressive rock groups | Psychedelic rock music groups | 1960s music groups | 1970s music groups | 1980s music groups | 1990s music groups | Musical quartets | Rock and Roll Hall of Fame inductees | Grammy Award winners | Capitol Records artists | Musical quintets | Musical trios | Rock music duos | English rock music groups | Musical groups from London

Pink Floyd were an English rock band who achieved worldwide success with their psychedelic and progressive rock music. Their work is marked by the use of philosophical lyrics, sonic experimentation, innovative album art, and elaborate live shows. Pink Floyd are one of the most critically acclaimed and commercially successful rock music groups of all time. It is estimated that they have sold over 200 million

Origin	= London, England
Origin	= London
Alias	= The Tea Set, The Pink Floyd Sound, The Pink Floyd
Genre	= Progressive rock, Psychedelic rock
Labels	= EMI, Harvest, Capitol, Tower, Columbia

Pink Floyd

Pink Floyd in 1968 (from left to right): Nick Mason, Syd Barrett, David Gilmour (seated in front), Roger Waters, Richard Wright

Background information

Also known as	The Tea Set, The Pink Floyd Sound, The Pink Floyd
Origin	London, England
Genres	Progressive rock Psychedelic rock
Years active	1965–1996, 2005
Labels	EMI, Harvest, Capitol, Tower, Columbia
Associated acts	Sigma 6, Joker's Wild, Zee
Website	www.pinkfloyd.co.uk www.pinkfloyd.com

Former members

Syd Barrett
David Gilmour
Bob Klose
Nick Mason
Roger Waters
Richard Wright

✓ ✗

Fig. 2. *WiGipedia* components: (a) shows a wiki article for "Pink Floyd" with an embedded contextual graph of semantically linked articles. In (b) dark blue nodes represent the starting point for the graph generation. In this case three English bands: "Pink Floyd", "Led Zeppelin", and "Deep Purple". The remainder of the graph represents commonalities between them: light blue nodes are other wiki articles (i.e. "England") and tan nodes are Wikipedia categories (i.e. "1980s music groups"). A user created edge is highlighted in green. (c) shows the drop-down menu with a list of suggested labels for the new edge. In this example "Pink Floyd" and "England" are linked by "Origin". (d) shows a confirmation step that occurs before the new update is propagated to Wikipedia.

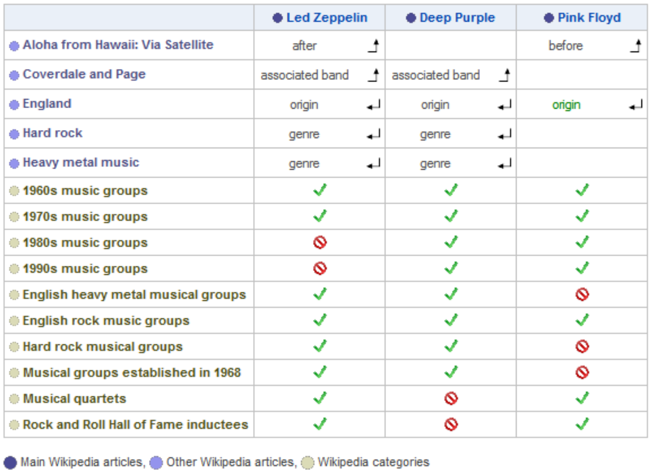
node have different relationship types, but originate from the same type/class of nodes. Note: this is usually a sign of inconsistency, but not a guarantee of inconsistency. For example, in Wikipedia the bank entities "Bank of America" and "Wells Fargo" are connected to the article "Financial Services" through an infobox property "industry". However, the bank entity "Wachovia" is connected to "Financial Services" through an infobox property "genre", and "Citibank" is connected via "products".

5) *Missing links*: When relations have not been provided in the Wiki. In the above example, three of the four banks are connected to the "United States" article. However, there is no direct link between "Citibank" and "United States", even though Citibank is a U.S. company.

In this paper, we are primarily focused on the latter two classes of inconsistency. Since relational queries issued over inconsistent semantic data usually return incorrect/incomplete results, we believe that the approach used by *WiGipedia* is a useful contribution to the Semantic Web community as it can improve such consistencies and therefore performance of relational queries over structured wiki data in the longer term.

In summary, the main aims of the *WiGipedia* system can be categorized as follows:

1) *Reduce Incompleteness* - Provides users with a mechanism to detect and create missing article links.
2) *Increase Consistency* - Promotes consistency and accuracy of structured data spread accross multiple articles.
3) *Facilitate Wikipedia Editing* - Provides users with an intuitive interface that allows Wikipedia edits in a few clicks, without requiring knowledge of the Wikipedia



	● Led Zeppelin	● Deep Purple	● Pink Floyd
● Aloha from Hawaii: Via Satellite	after ⤴		before ⤴
● Coverdale and Page	associated band ⤴	associated band ⤴	
● England	origin ↵	origin ↵	origin ↵
● Hard rock	genre ↵	genre ↵	
● Heavy metal music	genre ↵	genre ↵	
● 1960s music groups	✓	✓	✓
● 1970s music groups	✓	✓	✓
● 1980s music groups	⊘	✓	✓
● 1990s music groups	⊘	✓	✓
● English heavy metal musical groups	✓	✓	⊘
● English rock music groups	✓	✓	✓
● Hard rock musical groups	✓	✓	⊘
● Musical groups established in 1968	✓	✓	⊘
● Musical quartets	✓	⊘	✓
● Rock and Roll Hall of Fame inductees	✓	⊘	✓

● Main Wikipedia articles, ● Other Wikipedia articles, ● Wikipedia categories

Fig. 3. Tabular representation of the same data from Figure 2. User input data is once again highlighted in green.

markup language or templates, etc.

4) *Provide Context* - Enriches the wiki with relevant contextual information per article and reveals connections between multiple related articles.

II. Usage Scenario

Figure 2 highlights the steps in the process flow of *WiGipedia*. For Bob, a casual Wikipedia user, the interaction experience occurs as follows: Bob searches Wikipedia for an article of interest, for example, Pink Floyd. The article page shows the standard wiki page and an infobox containing a picture with some facts about the band. In addition to the infobox, Bob notices a graph with nodes and edges embedded in the wiki page. The graph contains nodes representing Pink Floyd and a range of other contextually relevant information such as music genres, places of origin, years of performance, and a selection of similar bands, e.g. Led Zeppelin and Deep Purple. Bob highlights a few nodes and notices that he can move them around to reconfigure the entire graph layout into meaningful arrangements which highlight important information. When he is satisfied with his layout, Bob then notices that all three bands on the periphery of the graph are linked to the node "English rock music groups", but only two of them are linked to "England". Bob decides to create a link between the nodes "Pink Floyd" and "England" and a suggestion box appears above the graph. The box contains a drop-down list of recommendations for the edge label. Bob notices that the other two bands, Led Zeppelin and Deep Purple, are linked to England via edges labeled "origin" so instinctively he picks the same label from the drop-down list. Then, the new labeled edge appears on the graph. Optionally, Bob can view or edit the same data by toggling to the tabular interface shown in Figure 3. Bob also notices that the Wikipedia article infobox has now changed, and that the text "England" has appeared and is highlighted in yellow, alongside a green check mark and a red X. Bob clicks on the green check mark and confirms his Wikipedia update. In a similar fashion, Bob can remove or update existing edges.

III. Related Work

WiGipedia combines facets from semantic web research, focusing on the gathering of rich semantic data in a collaborative manner, and on visualization and editing of Wiki data A discussion of the state of the art in each of these research areas follows.

A. Wikipedia and the Semantic Web

The method presented in this paper requires that complex queries can be issued over structured Wikipedia data. For example, such a query might be "show me similarities between Led Zeppelin, Pink Floyd, and Deep Purple". DBpedia is an online resource containing RDF representations of Wikipedia data, queryable in SPARQL. The DBpedia results for our sample query are shown in the visualization in Figure 2. We acknowledge that several comparable data sources exist, such as Freebase [5]. The deciding factor for developing this prototype to use DBpedia is that it essentially mines only from Wikipedia, providing a tightly-scoped closed-ecology for the content. Continuing work on this tool will examine the relative merits of both DBpedia and Freebase for the purpose of "cleaning" inconsistent data in Wikipedia.

There are many other tools which harness structured DBpedia data for various applications. A good representative example is Vispedia [6]– a system that provides user-compiled visualizations of Wikipedia data, which is integrated with the standard wiki page through hyperlinks. Lee et al. present a similar technique in [7] for combining semantic similarity with traditional text based search to improve search results. Similarly, in our approach, we begin with text based search (seed articles), and show semantic linkages to other articles of potential interest to a user. RelFinder (RF) [8], [9], similiarly to *WiGipedia* is a tool that supports relational queries over Wikipedia and reveals relationships within a set of known objects by displaying its results as a graph visualization. RF, however, is not used as an input modality to improve the data.

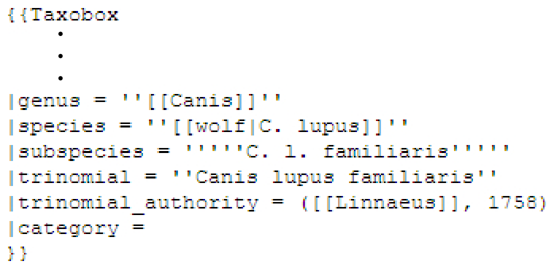
B. Visualization of Wikipedia

In addition to many efforts to produce static visualizations of Wikipedia data [10] to date, there have also been many efforts to create visualizations that communicate various dynamic aspects of the 3.23 million[2] articles in the English Wikipedia, such as [11], [12]. The Cimple project [11] developed a web framework focusing on visual analysis of the underlying social network formed around edits on various articles. Vispedia also employs visualization for data exploration and sense-making based on a user's information gathering requirements. While *WiGipedia* does provide contextual information about articles through visualization, it contrasts with the Vispedia approach in that *WiGipedia* uses a interactive interface facilitating *input* of information, similar to approaches by Hoffmann [13] and Krieger [14].

C. Wikipedia Editing Tools

Autonomous systems such as Kylin [2] train machine learning algorithms over Wikipedia infoboxes. Wu et el. estimate that 10M new facts can be added to DBpedia through the Kylin algorithms. While this is clearly a valuable contribution, we believe that it remains important to maintain a level of human supervision and input to foster healthy growth in Wikipedia. KOG [3] takes the Kylin approach one step further and aims to automatically refine the structure of the infobox ontology through SVMs and joint inference techniques. It is shown in [3] how this can lead to improved querying in Wikipedia. Holloway et al. [15] present a study of the semantic graph of Wikipedia and find that the occurrence of categories in articles follows a power law distribution and can reveal a clustered graph of semantic structure. However, as Wu et al. correctly argue, the existing category system in Wikipedia can be noisy, redundant and incomplete. Our approach to "cleaning" semantic data in Wikipedia focuses on linking articles to

[2] http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia



```
{{Taxobox
    .
    .
    .
|genus = ''[[Canis]]''
|species = ''[[wolf|C. lupus]]''
|subspecies = '''''C. l. familiaris'''''
|trinomial = ''Canis lupus familiaris''
|trinomial_authority = ([[Linnaeus]], 1758)
|category =
}}
```

Fig. 4. Example of the existing infobox markup in Wikipedia of the main article about the United States

other articles or categories through collaborative user input via simple to use tools. While we believe that this approach is capable of gathering sufficient data to address noise in article-category memberships, we note that it will not have an effect on the existing hierarchical structure of Wikipedia categories. In contrast to other semantic wiki efforts, *WiGipedia* attempts to allow users to 'close the loop' between DBpedia and Wikipedia through visualization and interaction, as illustrated in Figure 1.

Figure 4 shows an example of the current MediaWiki mark up for an infobox. While a computer scientist might comment that it is simply a parenthesized list of key-value pairs, and is very easy to understand, this does not hold for an average user. We posit that a small-scale image of few connected entities, either as a graph or table will be easier and faster for an average user to understand. Studies such as the MERIT [16] and Wikimedia Usability Initiative [17] clearly highlight that average users have difficulty editing Wikipedia. Accordingly, there have been several efforts aimed at providing simple editing tools for Wikipedia. For example, WikiHow [18] supports guided editing where the user is walked through a wizard-like process. WikiHow reports that 30% of editors continue on to push save. [1]. Wikia is another wiki that contains a WYSIWYG editor called CKeditor (also available as an extension for Wikipedia, if a user knows how to install it). It provides toolbars for text editing, images, infobox templates and various other intuitive controls. Wikia report that 50% of people who begin editing finish and press save [1]. However, the Wikia editor does not support more complex edits such as tables and some infoboxes, and requires switching to source-mode to complete in these cases. In this paper, we are interested in a particular specialization of the editing process, in that we focus on linking entities contained in infoboxes and other structured components. This contrasts with existing editing tools in that its focus is simpler, more task-specific, and comprised of potentially single click use cases.

IV. Design and Implementation

The goal of the interface is to explain how data in the current Wikipedia article relates to similar data in related articles as exploration tool, and allow for discovery and correction of inconsistencies as contribution interface. *WiGipedia* uses both a node-link graph and a traditional tabular view which can be toggled to at the user's discretion. Both representations visualize the same data and provide equivalent editing functionality. We treat the graphical mode as primary and default for several reasons. It offers a more engaging visual, contained in a more concise footprint for the sake of supplementing an article's summary section with some orienting context and overview. Graphs can provide more of a direct feeling for multi-hop relationships than tables can, which is especially important for its role of eliciting refining edits. The tabular view is provided to strengthen the observability of side-by-side type comparisons as well as to provide an alternative for users uncomfortable with graphs.

A. Data Generation

We generate a semantic graph in which every node corresponds to a Wikipedia article and every edge corresponds to a semantic relationship between two articles extracted from the wiki infoboxes. In terms of DBpedia, every node-edge-node triple corresponds to a subject-predicate-object RDF triple. All edges are directed and labeled with the RDF predicate term to indicate the type of relationship they express.

To start generating the graph, we take the current Wikipedia article the user is viewing along with a selection of related articles collected by one of a variety of means discussed below. These articles form what we call the "principal set", which are the main subject for comparison. SPARQL queries on DBpedia[3] are used to obtain additional articles linked to at least two of the principal set articles. The queries default to including direct and 1 hop links, and can be configured to include 2 hops. However, as warned in [8], queries of any more than 2 hops away become restrictively slow. The Web Service-style HTTP requests for the queries are spawned in parallel when possible to improve performance.

B. Graph View

The graph interface is based on a modified form of the existing WiGis framework [19] for the purposes of rendering an interactive web-based graph across browsers without the need for any plugins. The visualized graph in *WiGipedia* is comprised of two tiers of nodes: the principal set nodes are arranged around a circular perimeter, and the remaining nodes populate the interior (Figure 2(b)). The principal set nodes are rendered in dark blue and always represent Wikipedia articles. The sub-nodes can be either categories (rendered in tan), or article pages themselves (rendered in light blue).

To layout the interior, the principal nodes are pinned around an outer circle and then a simple Fruchterman-Reingold force directed algorithm is applied [20].

To help the user make the connection between the types of graph nodes and their corresponding article elements, a circle of matching color is placed next to the article title, infobox, and category list (Figure 2 (a)).

[3] http://dbpedia-live.openlinksw.com/sparql/

C. Table View

The Pathetic Fallacy, as applied to RDF graphs [21], warns about the assumption that just because the internal data structures are graphs, a matching graph-based user interface should be used to view and interact with them. The mainstream interfaces used for viewing and editing structured data are based on tables and forms. Freebase for example, has an entirely form-oriented interface, and when it comes to Mediawikis, the Semantic Forms[4] plug-in is the standard for editing. Google Refine[5] has particular overlap with the goals of our interface in that it is designed to facilitate cleansing data and improving consistency across a data set. The *WiGipedia* tabular view is designed to portray the exact same data set as the graph (Figure 3). We orient the principal set of articles as the columns and all other referenced articles as the rows whether they appear as the subject or the object in the RDF triples. We indicate the direction next to the predicate in the individual cells using an arrow icon which either points left to the row heading or up to the column heading as the object of the triple. Arrow direction can be flipped by clicking directly on the icon. A table has an easier time visually scaling to large amounts of data without becoming cluttered as the graph would.

D. Generation of the Principal Set

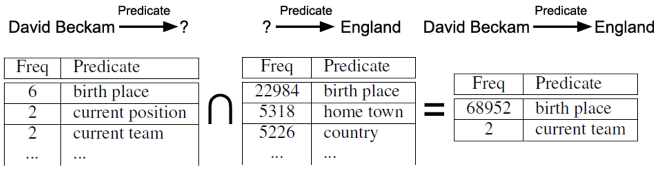
The user can input a principal set of articles to generate a graph they are interested in seeing using the *WiGipedia* toolbar (found above the wiki article in Figure 2(a)), but we generate a set automatically to provide an engaging initial context that is still relevant and useful. Initially, we generate a list of the page's outgoing hyperlinks sorted by occurrences (Table I). Then, links not in the top twenty percent are removed.

Principal set nodes are picked randomly with more weight given to articles of the same type as the current. Having articles of the same type is helpful because they share the same schemas which in turn create lower-hanging-fruit edits for the user. The user study discussed below reaffirmed this as we noticed that users made substantially more edits when the principal set articles were of the same type. As exemplified by Figure 3 users can easily spot the trend that music bands are linked to countries by the "origin" property and create a new link with higher confidence.

1) Alternative Methods: We have considered alternative methods for generating the principal article set. One of which uses a history of recently browsed Wikipedia pages. The benefit is not requiring mining and increasing the likelihood of relevance/familiarity to the user. However, an issue is the lack of guarantee of article relation. As another approach, we could take the top few nodes with the greatest connectedness to the target from within DBpedia. This would offer a more comprehensive structured data oriented metric which considers incoming links as well. A third approach is to base it solely on category membership, fetching other popular items in the various categories the current article belongs.

Hits	Article	Hits	Article
26	United_States	8	Blues
11	Bebop	8	Jazz_fusion
9	Free_jazz	8	Dixieland
9	Hard_bop	6	Swing_music
9	Miles_Davis	6	Africa

TABLE I
EXAMPLE COLLECTION OF THE 10 MOST REFERENCED PAGES (WITH OCCURRENCE COUNTS) AS SCRAPED FROM THE JAZZ ARTICLE ON WIKIPEDIA. NOTE THE DIVERSITY OF 'DATA TYPES' - PLACES, GENRES, PERSONS, AND OBJECTS.



Freq	Predicate
6	birth place
2	current position
2	current team
...	...

Freq	Predicate
22984	birth place
5318	home town
5226	country
...	...

Freq	Predicate
68952	birth place
2	current team

Fig. 5. Generating the DBpedia edge label suggestion set

E. Recommendation and Ranking of Edge Labels

When the user performs an edit, four different combinations of node type pairs can be selected, only three of which are valid: article–>article, article–>category, category–>category. The pair category–>article is not semantically supported by Wikipedia. The addition of an edge between two articles requires the user to provide a label to indicate the intended relation type. We assist the user in deciding on a label not only to be helpful but also to facilitate consistency and coherence within the existing Wikipedia corpus. We aggregate edge label suggestions from three different interdependent sources: the displayed semantic graph, an additional DBpedia query, and the source article's wiki infobox template.

- *Graph*: The graph is what users see, and is likely to be the scope of the considerations used to inspire its own modification. For a precedent-based angle, we take the union of the set of outgoing edge labels from the source node and the set of incoming edge labels to the destination node (Figure 2 (b) edges in red).
- *DBpedia*: For a semantic compatibility based angle, an additional query is made to DBpedia which looks at the intersection of the same two sets, but this time it is comprehensive across all RDF triples for the source and destination nodes, not just what makes it into the graph. By taking this intersection we ensure that the proposed edge labels make semantic sense for both outgoing source node edges and incoming destination node edges (Figure 5).
- *Wikipedia*: Wikipedia infoboxes are usually associated with a template type which dictates a loosely standardized superset of attribute names which have been used commonly among its instances for similar articles (e.g. http://en.wikipedia.org/wiki/Template:Infobox_artist). These names are used for label suggestion. To ensure robustness in the rare cases when the infobox does

[4] http://www.mediawiki.org/wiki/Extension:Semantic_Forms

[5] https://code.google.com/p/google-refine/

Component	T (s)	% Total
Load a Wikipedia page	1.5	4.7
Generate principal set	0.3	0.9
Query DBpedia for direct and 1-hop away links	18.3 (0.3)	57.7
Generate semantic graph	0.5	1.6
Embed visualization in the Wikipedia page	1.6	5.0
Choose two article nodes and send a link request	0.5	1.6
Generate label suggestion list for the new edge		
From the graph	0.2	0.6
From DBpedia	5.1 (0.1)	16.1
From Wikipedia	1.6	5.0
Preview modified Wikipedia article and confirm	2.1	6.0
TOTAL	31.7 (8.7)	100

TABLE II
COMPUTATION TIMES FOR LINKING TWO WIKIPEDIA ARTICLES THROUGH THE *WiGipedia* INTERFACE. WHEN DATA IS LOADED FROM CACHE INSTEAD OF DBPEDIA THE TIMES ARE SHOWN IN PARENTHESES.

not refer to a template, we combine the field names embedded in the raw article text with those from the template. This also helps to catch the tags which may be unique or specific to the article, and therefore potentially of special interest to the user when adding their new edge.

We pool and rank the output from all of these sources together into a drop down list for selection by the user (Figure 2 (c)). Each entry is ranked by the number of sources it appeared in. In the case of a tie, an entry is chosen at random. If no suggestion will suffice, the user is always welcome to use the edit box to provide their own label.

V. EVALUATION

A. Computation Time

For computation time performance evaluation, end-to-end work flow stages were automated and times recorded. The results are listed in the left column of Table II. The computation times and percentages shown on the right reflect the averages for each step over 1000 trials using our university network. The dominant stage is the SPARQL query to DBpedia; sticking to no more than 1 hop relations and 3-4 principal nodes, it averages over half the total process time. Depending on the structure of the linked data, some queries to DBpedia are issued in parallel, while others need to be sequential.

To address the delay caused by querying DBpedia (18.3 + 5.1 seconds), we cache the query results on the server that runs *WiGipedia*. Loading time for data once it has been cached are included in parentheses.

B. User Study

A controlled user study of 25 participants was performed to address the research questions discussed in the introductory text: can an interactive graph approach be used to gather semantic data from Wikipedia users quickly and easily? Do users prefer the *WiGipedia* interface to the current Wikipedia method for updating infoboxes? In addition to these central questions, the user study was designed to gain information about users' overall experience while using the system to learn contextual information about an article, and to provide semantic updates to an article. Specifically, the study looked at ease-of-use, and satisfaction levels from a qualitative perspective. In addition, the study examined the degree of latent information gain that occurred while users interacted with the visual interface. A comparison was made over a range of tasks between the current wiki update mechanism, and the *WiGipedia* interface.

The study opened with a pre-test questionnaire, followed by a familiarization task. Participants were then asked to perform two tasks: use the tool to add/modify/remove links between Wikipedia articles, and to assess the quality of edge label recommendations for the links they provided. The study concluded with a post-test questionnaire. On average, studies lasted 40 minutes.

The group of participants consisted of 17 males and 8 females, ranging in age from 19 to 43 with an average of 27 years. Figure 6 presents results on participants prior experience with Wikipedia and related technologies. On average, participants reported that they were familiar (2.96 on the Likert scale) with the way Wikipedia edits work. This is an unusually high value, since the WikiMedia foundation report that less than 8% of Wikipedia users have performed edits in the past [1]. This result is likely due to the fact that a third of the participants were graduate students in computer science and related fields and 5 of them have edited Wikipedia in the past. Most users were not familiar with the semantic web and RDF (1.96 on average), and even less familiar with DBpedia (1.57 on average).

1) Task 1: Sense-making and semantic linking: To evaluate the interactive graph as a sense-making interface and an input tool, participants were told that they could add, delete, or modify links on the graph based on their comprehension of the underlying semantic links. An average of 2.33 semantic links were added to each graph in this task, 0.24 links were modified, and 0.1 links were removed.

Participants were asked to select three to four articles to form the principal set. Under "normal" operation, a graph is embedded in each wiki page and the principal set is automatically selected through existing article links. For this task, articles were manually selected to ensure that the participant had a level of knowledge about the articles that enabled them to judge consistency of the connected data on the graph. Diverse topics were chosen by the participants, such as Music Composers, American Banks, Car Manufacturers, Tennis Equipment, Beer Brands, World War II Cities, European Ports, Italian Food, etc. Interestingly, there were cases in which users were able to make links about data that they knew nothing about prior to the study. For example a user noticed that American singer "Britney Spears" is listed in the category "Baptists from the United States" but not in the category "American Christians". While it is generally possible to infer such information though automated algorithms, we posit that by allowing real people to examine link structures in a clear, easily accessible manner, we can employ a wisdom of crowds approach to aggregation of data which will eventually catch errors in pre-existing link structures that an automated algorithm will not detect.

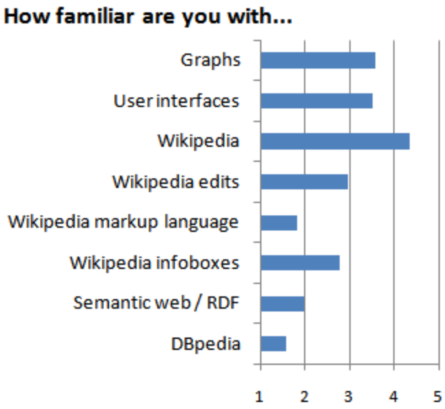


Fig. 6. Pre-study questionnaire results

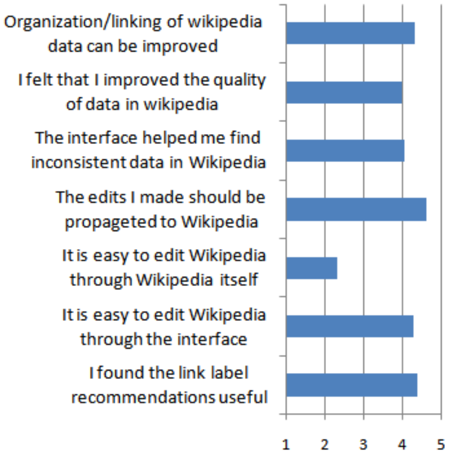


Fig. 7. Post-study questionnaire results

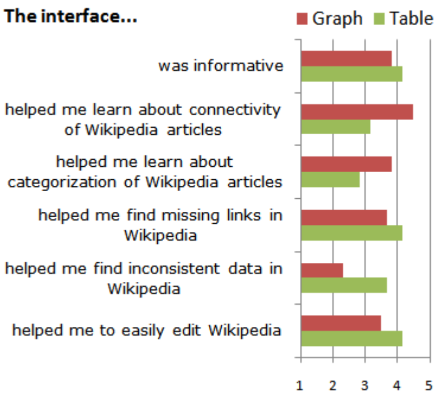


Fig. 8. Interface questionnaire results

2) Task 2: Link label recommendation: Section IV-E described the three sources used by *WiGipedia* for prediction of edge labels. To evaluate the usefulness of each of the three link recommendation sources, we again asked study participants to use the graph to add a link between two Wikipedia articles (blue nodes in Figure 2 (b)). 23 of the 25 participants were able to find and create such new relation in their graphs. Then, participants were asked to write down a label for the relation, such as Pink Floyd relates to England by “origin”, for instance. This approach was used in an effort to gather opinions that were unbiased by the system’s suggestions. After providing a label, participants were then presented with a ranked list of suggestions in a drop-down menu, as illustrated earlier in Figure 2(c). If a participant felt that one of the system’s label predictions was more accurate for the new relation, they selected it via a mouse click, thereby discarding their original label.

For each link that participants created, we recorded the percentage of times that each recommendation source contained the edge label that was eventually chosen by the participant (Table III). The graph-based prediction technique was 100% accurate on this metric, that is, the list of recommendations generated by the graph always contained some satisfactory label for the new connection. Given that the average size of the recommendation list generated by the graph was 3, this is a promising result. DBpedia produced the smallest sized recommendation set at 1.1 items on average, out of these lists, about half of them were chosen by the participant. This indicates that taking the intersection of outgoing predicates from a given subject with incoming predicates into a given object across all relevant DBpedia RDF triples can produce reasonably reliable suggestions for semantic links. The Wikipedia set contained 18.7 items on average and was 73% successful. Out of the 23 participants who completed the task and were able to make updates through the system, 20 of them chose the top ranked link in the suggestion drop-down box.

Technique	Average Set Size	Percent Picked
Graph	3	100%
DBpedia	1.1	45%
Wikipedia	18.7	73%

TABLE III
COMPARISON OF PERFORMANCE OF EDGE LABEL RECOMMENDATION TECHNIQUES. THE PERCENTAGE COLUMN SHOWS THE FREQUENCY THE CANDIDATE SET GENERATED BY A TECHNIQUE CONTAINED THE LABEL THAT THE PARTICIPANT SELECTED AS THE NAME FOR THE NEW LINK.

3) Task 3: Comparison with the standard Wikipedia editing mechanism: To gain some insight into the benefit of our approach to editing structured Wikipedia data, participants were asked to make a range of updates using both the *WiGipedia* interface and the standard Wiki interface. Time taken to input each update was recorded. First, we would like to note that a fair comparison between these interfaces is not easy to define. Wikipedia markup, as exemplified in Figure 4, is designed for editing the entire document, and we are focusing only on structured content. Also, aside from considering time taken to make updates, it is important to note that 5 users in this study reported that they were already familiar with Wiki markup (see Figure 6), and had edited Wikipedia in the past.

Results from our post-study questionnaire in Figure 7, show that all participants found the editing process substantially easier through the *WiGipedia* interface. Interestingly, out of the 20 users with no previous experience editing Wikipedia, 12 of them could not complete the infobox update task through the standard interface.

4) Graph vs. table interface evaluation: Having measured the ratio of time spent by each user on each mode of visualization, we were surprised at how even they came out to be. Users chose to look at the graph 53% of the overall time, while making 58% of their edits using the table’s input interface. The post-task survey indicates higher ratings of the graph interface only when it comes down to the connectivity and categorization questions, even though the overall ratings for both are reasonably high (Figure 8).

We additionally asked users why they used one or the other to make their edits. Of those who preferred the table, responses hinged on it being "more intuitive" and efficient, while those who preferred using the graph, hinged on a greater sense of control and satisfaction. "It is more satisfying to connect things like a web. I could see myself building it, creating something, making the connection." In some cases a cluttered looking instance of the graph turned the users away, even when they had hopes for a better layout, and were not often willing to attempt a better layout manually.

VI. Conclusion and Future Work

This paper introduced *WiGipedia*, a novel system that facilitates the input of semantic information from regular Wikipedia users. By querying DBpedia for semantic relations among a selected set of articles and generating graph- and table-based visualizations, the system provides context to the article being read. A main contribution of *WiGipedia* is as an input modality, supporting near single-click semantic updates to Wikipedia, based on the users' comprehension of relations in the interface. To make this process easy for the user, *WiGipedia* provides dynamic recommendations for link types. Results of a comparative user study indicate that users readily understand the interface and can provide semantic updates in less time than is necessary for the existing Wikipedia interface.

There are many open questions and areas for further research on *WiGipedia*. For example: How do we help users make better decisions about semantic relations? Can we improve recommendations of link types? On a larger scale, what impact will a 'wisdom of crowds' approach to data aggregation have on the quality of semantic updates? How will this vary with respect to the difficulty level of an article? Moreover, techniques such as Kylin [2] or Kog [3] can automatically predict new links on the embedded graph, and users can verify them. During our user study, a suggestion was made to implement *WiGipedia* as a stand-alone game, wherein two or more users have to generate matching semantic relations. Google have successfully applied this technique to the problem of image labeling[6].

Looking forward, we believe that contributing to the consistency and structure of Wikipedia is a step in the right direction towards the creation of a rich Wikipedia ontology, capable of supporting complex analytical queries over this huge knowledge repository.

VII. Acknowledgment

This work was partially supported by NSF grant IIS-1058132; by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory contract number FA8650-10-C-7060; by the U.S. Army Research Laboratory and the U.S. Army Research Office under MURI grant No. W911NF-09-1-0553; and by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of IARPA, AFRL, ARL, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] A. Lih, "Can wikipedia survive popular success and community decline?" 2010, http://andrewlih.com/media/20100222-wikipedia-survive.pdf.

[2] F. Wu and D. S. Weld, "Autonomously semantifying wikipedia," in *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. New York, NY, USA: ACM, 2007, pp. 41–50. [Online]. Available: http://portal.acm.org/citation.cfm?id=1321440.1321449

[3] ——, "Automatically refining the wikipedia infobox ontology," in *WWW '08: Proceeding of the 17th international conference on World Wide Web*. New York, NY, USA: ACM, 2008, pp. 635–644. [Online]. Available: http://portal.acm.org/citation.cfm?id=1367583

[4] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *Proceedings of 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007)*, November 2008, pp. 722–735. [Online]. Available: http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/Auer-Bizer-ISWC2007-DBpedia.pdf

[5] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 2008, pp. 1247–1250. [Online]. Available: http://ids.snu.ac.kr/w/images/9/98/SC17.pdf

[6] B. Chan, L. Wu, J. Talbot, M. Cammarano, and P. Hanrahan, "Vispedia: Interactive visual exploration of wikipedia data via search-based integration." *IEEE Trans. Vis. Comput. Graph.*, vol. 14, no. 6, pp. 1213–1220, 2008.

[7] M. Lee, W. Kim, and T. G. Wang, "An explorative association-based search for the semantic web," in *Proceedings of the 2010 IEEE Fourth International Conference on Semantic Computing*, ser. ICSC '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 206–211. [Online]. Available: http://dx.doi.org/10.1109/ICSC.2010.17

[8] P. Heim, S. Hellmann, J. Lehmann, S. Lohmann, and T. Stegemann, "Relfinder: Revealing relationships in rdf knowledge bases." in *SAMT*, ser. Lecture Notes in Computer Science, T.-S. Chua, Y. Kompatsiaris, B. Mrialdo, W. Haas, G. Thallinger, and W. Bailer, Eds., vol. 5887. Springer, 2009, pp. 182–187.

[9] J. Lehmann, J. Schüppel, and S. Auer, "Discovering unknown connections - the dbpedia relationship finder," in *Proceedings of the 1st SABRE Conference on Social Semantic Web*, 2007.

[10] Chris harrison - wikiviz: Visualizing wikipedia. [Online]. Available: http://www.chrisharrison.net/projects/wikiviz/index.html

[11] A. Doan, R. Ramakrishnan, F. Chen, P. Derose, Y. Lee, R. Mccann, and M. Sayyadian, "Community information management," *IEEE Data Eng. Bull*, vol. 29, p. 2006, 2006.

[12] E. S. B. Pirolli, P. L.; Wollny, "So you know you're getting the best possible information: a tool that increases wikipedia credibility." in *27th Annual CHI Conference on Human Factors in Computing Systems (CHI 2009)*, 2009.

[13] R. Hoffmann, S. Amershi, K. Patel, F. Wu, J. Fogarty, and D. S. Weld, "Amplifying community content creation with mixed initiative information extraction," in *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*. New York, NY, USA: ACM, 2009, pp. 1849–1858.

[14] M. Krieger, E. M. Stark, and S. R. Klemmer, "Coordinating tasks on the commons: designing for personal goals, expertise and serendipity," in *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*. New York, NY, USA: ACM, 2009, pp. 1485–1494.

[15] T. Holloway, M. Bozicevic, and K. Börner, "Analyzing and visualizing the semantic coverage of wikipedia and its authors: Research articles," *Complex.*, vol. 12, no. 3, pp. 30–40, 2007.

[16] R. Glott, "Wikipedia survey," 2009.

[17] "Wikimedia usability initiative," 2010, http://usability.wikimedia.org.

[18] "Wikihow," 2010, http://www.wikihow.com.

[19] B. Gretarsson, S. Bostandjiev, J. O'Donovan, and T. Höllerer, "Wigis: A scalable framework for web-based interactive graph visualizations," in *GD'09: Proceedings of the International Symposium on Graph Drawing*, 2009.

[20] T. M. J. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software - Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, 1991. [Online]. Available: citeseer.ist.psu.edu/fruchterman91graph.html

[21] D. Karger and M. Schraefel, "The pathetic fallacy of rdf," 2006, position Paper for SWUI06.

[6] http://images.google.com/imagelabeler/