VisTA: Visual Terminology Alignment Tool for Factual Knowledge Aggregation

Anastasia Axaridou[1[0000-0002-7598-111X]], Konstantina Konsolaki[1[0000-0001-8002-3250]], Maria Theodoridou[1[0000-0002-4623-9186]], Artem Kozlov[2[0000-0002-8666-4347]], Peter Haase[2[0000-0002-7561-7000]] and Martin Doerr[1[0000-0002-6006-0844]]

[1] Institute of Computer Science, FORTH, Heraklion, Greece
{axaridou, konsolak, maria, martin}@ics.forth.gr
[2] metaphacts GmbH, Walldorf, Germany
{ak, ph}@metaphacts.com

Abstract. The alignment of terminologies can be considered as a kind of "translation" between two or more terminologies that aims to enhance the communication among people coming from different domains or expertise. In this paper we introduce the Visual Terminology Alignment tool (VisTA) that enables the exact alignment for RDF/SKOS-like terminologies, suitable for integrating knowledge bases, rather than information retrieval systems. The tool provides a simple and friendly web-based user interface for the alignment between two terminologies, while it visualizes the terminology hierarchies, enables the interactive alignment process, and presents the alignment result. The latter is a native RDF/SKOS graph that interconnects the two terminology graphs, supporting interoperability and extending the search capabilities over the integrated semantic graph, using the broader and exact match properties.

Keywords: Terminology Alignment Tool, Thesaurus, Visual, Interactive, Exact Matching

1 Introduction

The alignment of terminologies, also referred to as thesauri mapping, can be considered as a kind of "translation" of the terms between two or more terminologies in order to support the communication among individuals, teams and communities coming from different scientific, cultural, technical or other domains. What is actually achieved by alignment is the integration of knowledge towards the completeness of expressivity and a better understanding of the entities used in multiple fields of life and science. In our approach *Terminology*, is a SKOS[1]-like description of *universals* (classes, types, etc.) [1], rather than of individual items (e.g. persons and places), employing the SKOS properties of *broader* and *exact* match.

The process of alignment is not an easy work. Several terms may be used for a concept, as well as different concepts may be expressed by the same term. Further-

[1] https://www.w3.org/2004/02/skos/

more, actual generalization and specialization hierarchies frequently break the semantics, inserting narrower terms incompatible with the meaning of the broader term, such as *natural phenomena* appearing under *Activities* (see example in section 3). Things are getting worse when large vocabularies are involved and consequently those situations can be treated by approximate matching methods producing inaccurate alignment results that cannot be totally trusted. Such results always suffer from their level of precision and recall containing an amount of erroneous or missing matches that in the end can be resolved only by the human intervention. A deep analysis in [2] shows ways in which thesaurus creators can improve their methodology to meet the challenges of networked access of distributed collections.

In this paper we are proposing the support of the principles of concept-based [2] alignment with a Visual Terminology Alignment tool (for short VisTA). VisTA aims to help the users to work on the intellectual handling of the alignment between two terminologies based on two essential hierarchy relationships: *broader-narrower* and *exact match*. We don't deal with inexact match, which works well with information retrieval but not with logical queries, and does not preserve recall and generalization semantics required in knowledge bases. In need of an exact alignment solution, VisTA is ideal for the alignment of small to medium sized terminologies and may also scale up to larger cases when running on sufficiently powerful machines.

VisTA is a module that conforms to a respective component foreseen by the Synergy Reference Model[2], an initiative of the CIDOC CRM Special Interest Group[3], for manipulation of data provisioning and aggregation processes, aiming at defining the processes needed to be executed or maintained between a data provider (the source) and a data aggregator (the target) institution [3], such as a museum and Europeana[4].

In the next sections we present features and the interface of VisTA. In section 2 we discuss the motivation and related work. Section 3 introduces the alignment problem in VisTA. Section 4 describes the user interface and the features of VisTA in terms of visualization of the alignment process and the produced result. Section 5 presents the template mechanism that provides configurable application components. Section 6 concludes with a discussion on the tool's current state and future work.

2 Motivation and Related Work

In the last two decades a lot of work has been published in the broader area of alignment, comprising schema matching, schema mapping and ontology alignment, focusing on the development of approaches trying to achieve as accurate results as possible [4]. Several frameworks have been proposed to apply automatic [5] or semi-automatic [6] procedures based on configurable workflows in order to produce alignment, usually for large vocabularies. They sometimes provide interactive ways [7] to build the alignment workflows by combining predefined matching algorithms. These algorithms may have to be fed with input parameters in intermediate steps of the align-

[2] http://www.cidoc-crm.org/sites/default/files/SRM_v1.5.pdf

[3] http://network.icom.museum/cidoc/working-groups/crm-special-interest-group/

[4] https://www.europeana.eu/

ment process to finally produce an estimation of the similarity among terms. It’s true that large vocabularies cannot be easily handled manually by a user, so the development of batch production solutions seems to currently be the best approach. But in such cases, although the process accelerates, an inexact result may be produced requiring further editing in order to become practically exploitable.

Some ontology alignment tools propose interesting solutions for the visualization of the alignment [8]. The graphical presentation of correspondences with connection paths between the entities of the aligned graphs [9] and the graph node representation [10] are the typical solutions provided. These methods are helpful for viewing the details of the alignment structure but suffer from overload and confusion with the amount of interconnections displayed. Some of the alignment tools offer a web interface [11-14] while others operate as standalone applications. A disadvantage of the latter is that the user gets involved in the installation procedure, and sometimes that discourages working with a tool. In all the above cases, the visualization of the alignment regards the representation of the correspondences between terms, based on the accepted similarity values against a defined similarity threshold.

The need to align terminologies in the Cultural Heritage domain made us study the work done in the alignment space. We concluded that there isn’t enough substantial work dedicated to terminology alignment used in knowledge bases, since it is currently considered as an inseparable part of the alignment bulk. The YAM++ online tool[5] [11] and the Amalgame[6] framework aim at thesauri matching. The second helps the user to build the alignment workflow by configuring and combining the available matching algorithms using a graphical representation of the workflow. Both tools result in similarity estimation of the terms providing a graphical interface asking the user to apply the correct relationship for each pair of terms, but without graphical attribution of the actual meaning of their activity.

Table 1. General state of the available alignment methods

Method	**Supportive GUIs**	**Input size**	**Output accuracy**	**Exploitability of result**
Automatic	Yes, for configs	Large	Approximate, e.g. estimation of similarity between terms	?
Semi-Automatic	Yes, for configs. Sometimes also for the manual phase but without graphical attribution of the semantics			
		Medium/Small	Exact, i.e. define the exact relation between terms	Yes
Manual	VisTA. Other?			

Table 1 presents the general state of alignment w.r.t. the applied methods, the available GUIs, the input size, the output accuracy and the exploitability of the produced result. As shown in the table, related work can be found on automatic and semi-

[5] http://yamplusplus.lirmm.fr/

[6] http://semanticweb.cs.vu.nl/amalgame/

automatic solutions, where supportive GUIs help mainly for the configuration of the algorithms and the workflows applied to the alignment process. Nevertheless, the result they produce is of a semantically uncertain exploitability.

Our belief is that terminology alignment of RDF/SKOS-like vocabularies can be treated in more precise ways, especially when application requirements arise for using knowledge bases. We mention the lack of work on the *exact (accurate) terminology matching* for the cases that a “correct” [2], as a question to user convention, alignment is needed and we offer VisTA to the terminology authors for efficient easy alignment. VisTA supports a *target-driven* alignment with *reconciliation* of a source terminology and using the broader and exact-match, two fundamental hierarchy relationships, allows the users for working top-down, controlling and preserving the generalization/specialization semantics known as *is-a*.

This intentionally simplified approach is materialized on a web interface, where we tried to visualize the alignment process between two RDF/SKOS-like terminologies. The applied hierarchy relations can be pre-configured and substituted with equivalent RDF links from other schemas. The alignment result is a native RDF/SKOS-like graph consisting of the necessary data interconnecting the involved terminologies.

3 The alignment problem in VisTA

In VisTA we deal with the *target-driven* alignment of two terminologies as an *asymmetric* process aiming to the *subordination* of a source to a target terminology. The process results in *$n \times m$ correspondences* among the terms of the two terminologies, whereas a *correspondence* is regarded as a direct association between two terms.

For the successful subordination of the source and the production of consistent knowledge, we assume three *principles*:

- Terminologies and the alignment result are *acyclic* graphs preserving the taxonomy *subsumption (is-a)*
- The *broader* and *narrow* relations are *symmetrically inversed*
- *Reconciliation of the source terminology* with the target but not the reverse

The alignment process results in the integration of different terminologies, under one target terminology which is considered as the core structure. The reconciliation of the structure of the source terminology is supported by the tool, in order to remove any subterms of the aligned source term that violate the preservation of subsumption. An example is presented below. The extension of the target terminology is based on *broader match and* the *exact match* relations see section 4.2.

During the process *multiple-inheritance* of terms, i.e. a term can have multiple parents, and subhierarchy *overlaps* may occur. These situations are allowed unless they break subsumption. The constraints (rules) of the process are presented in section 4.3.

The alignment result (see section 4.4) really empowers the searching capabilities in a semantic network, as the users of different terminologies are enabled to make queries using the common target vocabulary together with their own familiar vocabularies, to find more resources in their results.

A Use Case Example. Suppose we are building a knowledge aggregator based on the generic *Backbone Thesaurus*[7] (*BBT*) and we need to integrate data from a provider that uses the *Art & Architecture Thesaurus*[8] (*AAT*). Using the BBT as the target terminology and the AAT as the source for the alignment process, the concepts of the AAT get subordinated to BBT which's structure is persisted. Both terminologies contain a facet named *activities*: in BBT, the facet classifies intentional actions, while in the AAT it encompasses areas of endeavor, physical and mental actions, etc. But the AAT *activities* facet includes in the taxonomy the concept of *natural phenomena* (Fig. 1). We argue that the latter violates the subsumption in BBT according to the BBT definition of *activities*. Hence, in order to set up an *exact-match* correspondence between the *activities* terms, it is important to exclude the inconsistent subhierarchies of the AAT term.

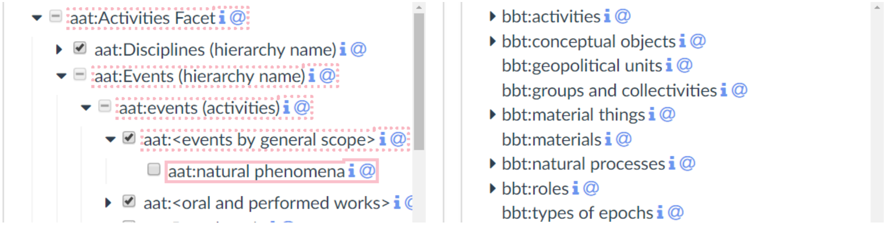

Fig. 1. The AAT term *natural phenomena* and the BBT hierarchy

4 VisTA Overview

4.1 User Interface

An easy to use interface with two tabs on the top of the main page takes the user to the respective working area: the Import/Export page, to manipulate terminology and alignment graphs and the Edit page, to proceed with an alignment process.

The Import/Export page provides the appropriate options to manage terminology and alignment data. Terminology data comprise distinct named graphs of SKOS/RDF data that can be used as source and target terminologies in an alignment process. The alignment data is a set of named graphs, the results of the alignment processes, each one related to a specific pair of terminologies. Helpful utilities are available for importing new terminology data and exporting the RDF graphs in multiple formats.

The Edit page provides appropriate components and functionality for performing the alignment steps as a directional process that takes place from the source to the target terms. For that purpose the page is divided in two main symmetric areas, the left one exposing the source terminology interface and accordingly the right one for exposing the target terminology.

[7] https://vocabs.dariah.eu/bbt/ConceptScheme/Backbone_Thesaurus/

[8] http://www.getty.edu/research/tools/vocabularies/aat/

An alignment process comprises drag'n'drop steps for the creation of distinct correspondences. On each step the currently aligned pair of terms is checked against the *alignment rules* (see 4.3). A dialog popup asks the user to choose which relation must be applied, prompting to *align a source to a target term as: exact or narrow term*. After the relation is selected by the user, the aligned source term appears on the target tree as a new tree-node or as a label extension to the target node (see section 4.2), depending on the selected relation, narrow or exact respectively.

4.2 VisTA Features

When developing an authoring tool, the design of proper functionality to help the user to work efficiently should be of high consideration. The specific challenge is to enable the user to maintain a mental image, an overview, of the process and progress of the alignment task, without being overwhelmed by the details or lost in details. VisTA offers a convenient and friendly interface to support *interactive* alignment, enabling the users to take charge of the alignment process and manipulate the data involved in the alignment activity. The alignment is treated as a directional drag'n'drop process from a source to a target terminology extending the target hierarchy. Visualization of the alignment on the target terminology tree, appropriate highlighting of the terms' state and specific accessibility options are powerful features of VisTA. The user is fully aware of the state of the alignment activity enabled for viewing the aligned terms in both their original structures as well as in the integrated semantic hierarchy.

Visualization of a Term. The terminology trees used in an alignment process consist of terms each having one of the three states used for visualization:

- *Aligned (or explicitly aligned).* The state for the source and target terms that are directly participating in a correspondence in the context of the alignment process (Fig. 2. a, b). These terms are both included in the alignment graph.
- *Implicitly aligned***.** The state for the source terms considered as indirectly aligned when one of their parent terms is being explicitly aligned (Fig. 2. c). These terms are included in the alignment graph linked to their aligned parents.
- *Not aligned.* The state for the source and target terms not included in the alignment graph. (Appear in the default black color in Fig. 2)

Visualization of the Terminologies. The representation of terminologies as tree views enables accessibility, comparability, and association between hierarchical structures. Taking advantage of the *lazy load* implementation, the tree view hierarchy is produced with two SPARQL queries. The first is responsible to fetch a multi-rooted terminology tree based on specific RDF properties for the hierarchy relation between broader and narrower terms. The second is responsible for fetching the children of a node on expanding. The queries and their properties are all predefined in a main *component template*. After the tree structures are created the term-nodes can be expanded/collapsed, selected and dragged by the user for the specific purposes of the editing and browsing.

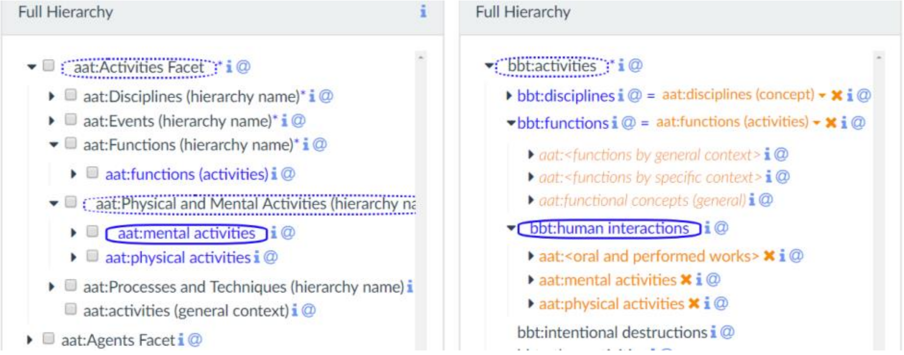
Full Hierarchy
aat:Activities Facet*
aat:Disciplines (hierarchy name)*
aat:Events (hierarchy name)*
aat:Functions (hierarchy name)*
aat:functions (activities)
aat:Physical and Mental Activities (hierarchy na
aat:mental activities
aat:physical activities
aat:Processes and Techniques (hierarchy name)
aat:activities (general context)
aat:Agents Facet
Full Hierarchy
bbt:activities*
bbt:disciplines = aat:disciplines (concept)
bbt:functions = aat:functions (activities)
aat:<functions by general context>
aat:<functions by specific context>
aat:functional concepts (general)
bbt:human interactions
aat:<oral and performed works>
aat:mental activities
aat:physical activities
bbt:intentional destructions

Fig. 2. a) Aligned terms are colored in blue in their original locations in the full hierarchies. b) Aligned source terms appear in orange at the target tree (right). c) The terms in italics (blue and orange), in both hierarchies, are considered as implicitly aligned. d) The blue circled terms visualize a correspondence e) The broader-narrow match is visualized with new orange tree nodes as children of a target term e.g. *aat:mental activities* is child to the *bbt:human interactions* f) The exact-match relation is visualized with the symbol "=" on a target node e.g. *bbt:functions = aat:functions (activities)* g) The exact-match terms share their children in the target tree e.g. *bbt:functions = aat:functions (activities)*

Creating and Visualizing Correspondences. An easy drag'n'drop mechanism from the source to the target terms helps the user to create correspondences (Fig. 2.d). The alignment result, as a set of correspondences, interconnects the aligned terminologies applying the two fundamental RDF properties of the SKOS schema:

- *skos:broader*, from a source to a target term in order for the source term to become a narrow term of the target, extending thus the *specification* of the latter, and
- *skos:exactMatch*, from a source to a target term, when the source term has an equivalent meaning to the target term, extending its *label* and *specification*

The semantics of the alignment are visualized, so the user is able to realize the results of this activity. In the *broader representation* the metaphor of a new branch as a new child of the target term, is used, carrying the source term sub-hierarchy from the original source tree. Thus an extended target tree is produced, with new branches coming from the source tree (Fig. 2.e). The user is allowed to edit the source term sub-hierarchy by excluding specific terms before creating a new correspondence.

In the *exact match representation* (Fig. 2.f) the aligned source terms do not appear as new branches of the target term but as a label extension of the target term joined to it with a "=" symbol next to the existing label. In this case the target node gets also extended by new branches coming from the sub-hierarchy of the source term. The first-level children of the aligned source term become new growing leaves and branches to the target node, implying this way that the equivalent terms can have the same sub-terms in common, as shown in Fig. 2.g.

The sub-terms of the aligned source term are considered implicitly aligned terms. VisTA distinguishes explicitly and implicitly aligned terms by specific highlighting on both the source and the target terminology trees. (Fig. 2. a, b, c)

Finally, an alignment correspondence can be removed from the target tree by deleting the specific explicitly aligned source term. Then, the term and its children are removed from the target tree and the correspondence triple is deleted from the graph.

4.3 Alignment Rules

During the alignment process, when a new correspondence is to be created between two specific terms, a check mechanism informs the user about the current state of the terms to prevent useless and inconsistent situations. The rules applied are:

- *Check for explicit (direct) alignment relation.* The source term must not be already aligned to the same target term, otherwise the process is canceled.
- *Check for the existence of the source term in the target tree.* When the source term to be aligned is already an *original* term of the target terminology then alignment of that term causes the *reconciliation* of the target terminology which is not allowed. In this case the user is informed and the process is canceled.
- *Check for implicit (indirect) alignment relation.* The source term may already be indirectly aligned to the target term. In this case the user gets a warning whereas the alignment is allowed.
- *Check for aligned descendants of the source term.* The source term may contain already aligned sub-terms. In this case the user gets a warning whereas the alignment is allowed.

4.4 The Result of the Alignment

What is actually produced by VisTA in an alignment process is a specific alignment graph. The format of that graph is *native* RDF/SKOS, and it contains all the explicitly produced information for the matching terms between the two terminologies as well as the children hierarchy of the terms coming from the source terminology. The structure of the aligned terms coming from the target terminology is considered to be constant, thus there is no need to copy their original structures into the alignment graph. Hence, to take advantage of the alignment result e.g. in a search operation, the appropriate searching space is the *union* of the alignment and the target terminology graphs.

5 Implementation and Configuration

VisTA is based on the template mechanism of the metaphacts platform[9]. The Metaphacts knowledge graph platform delivers an extensible template mechanism based on HTML5 Web Components[10] and Handlebars[11] templating engine. Out of the box

[9] http://metaphactory.com

[10] https://html.spec.whatwg.org/multipage/custom-elements.html#custom-elements/

the platform provides built-in configurable generic components for visualization, search and authoring of RDF knowledge graphs. In addition to built-in components the platform also offers Typescript/JavaScript based SDK that can be used to develop custom UI components or add additional functionality to already existing ones.

The main page of VisTA is the result of the rendering of a particular *component template*. The *parameters* included in the template can be externally modified to extend the operability to more terminology cases since the use of RDF relations cannot be foreseen in new terminologies. Some of the configurable parameters comprise: the definition of the RDF *broad-match* and *exact-match* relations used in the alignment process, the SPARQL queries templates for retrieving the *root* terms of a terminology tree, the *children* and the *parents* of a term in a terminology tree, the SPARQL query template for performing the *search* operation, the *RDF property paths* used in SPARQL queries templates for matching the *hierarchy relation* between a term-subterm and the *labeling* of the terms.

6 Conclusions

We presented VisTA, a tool that delivers an *interactive* solution for the exact terminology alignment problem, required in the context of data provisioning and aggregation processes. Terminologies and thesauri, contrary to other ontologies can be handled in more precise ways regarding alignment in order to produce accurate and exploitable results. VisTA offers a suitable and convenient web interface to visualize the alignment process between a source and a target RDF/SKOS-like terminology. The produced result is a native RDF/SKOS graph that interconnects the two terminologies towards the extension of searching capabilities over the integrated semantic graph.

VisTA has not been used on a regular basis yet. Several tests with medium sized terminologies used as sources to be aligned to the full AAT terminology, showed a good performance. VisTA, taking advantage of the *lazy load* tree implementation, has a fine response on rendering the tree hierarchies. On large vocabularies, there are delays observed when a new correspondence is being created (or deleted) related to performance issues on the select/update queries of the SPARQL endpoint.

Finally, we propose VisTA not as a competitive but as a complementary supportive solution to the work pending for the manual phase of an alignment procedure. Currently VisTA provides a from-scratch alignment process, but an extension of the tool is planned: to use as input, except from the pair of terminologies, furthermore a proposed base of the estimated similarity between terms (e.g. using the EDOAL[12] alignment format), as the basis for the manual process. Other features supportive to the users, planned for future versions, are: alternative views of correspondences or parts of the alignment graph, the extension to more than the two basic relationships, the extension of the alignment rules to other semantics.

[11] https://handlebarsjs.com/

[12] http://alignapi.gforge.inria.fr/edoal.html

Acknowledgments

This work is the result of collaboration between the British Museum, metaphacts and FORTH in the context of the ResearchSpace[13] project. We would like to thank all the RS team and especially Dominic Oldman for his kind support as well as Mike Kelly for his well-aimed remarks and suggestions.

References

1. Guarino, N., Formal Ontology in Information Systems, Amsterdam, The Netherlands: IOS Press, 1998.
2. Doerr, M., "Semantic Problems of Thesaurus Mapping," Journal of Digital Information, vol. 1(8), 2006.
3. Marketakis, Y., Minadakis, N., Kondylakis, H., Konsolaki, K., Samaritakis, G., Theodoridou, M., Flouris, G., Doerr, M., "X3ML Mapping Framework for Information Integration in Cultural Heritage and Beyond," International Journal on Digital Libraries (IJDL), vol. 18, no. 4, pp. 301-319, November 2017.
4. Shvaiko, P., Euzenat, J., "Ontology matching: state of the art and future challenges," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 1, pp. 158-176, 2013.
5. Noy, N.F., Musen, M.A., "The PROMPT suite: interactive tools for ontology merging and mapping," 2003.
6. Manakanatas, D., Plexousakis, D., "A Tool for Semi-Automated Semantic Schema Mapping: Design and Implementation," in International Workshop Data Integration and the Semantic Web, Luxembourg, 2006.
7. van Ossenbruggen, J., Hildebrand M., de Boer V., "Interactive vocabulary alignment," in Proceedings of the 15th international conference on Theory and practice of digital libraries: research and advanced technology for digital libraries, Berlin, Germany, 2011.
8. Granitzer, M., Sabol, V., Onn, K.W., Lukose, D., Tochtermann, K., "Ontology Alignment—A Survey with Focus on Visually Supported Semi-Automatic Techniques," Future Internet, vol. 2(3), pp. 238-258, 2010.
9. Aumueller, D., Do, H.-H., Massmann, S., Rahm, E., "Schema and ontology matching with COMA++," in Proceedings of the 2005 ACM SIGMOD international conference on Management of data, Baltimore, Maryland, 2005.
10. Lanzenberger, M., Sampson, J., "AlViz - A Tool for Visual Ontology Alignment," in Tenth International Conference on Information Visualisation, London, England, 2006.
11. Bellahsene, Z., Emonet, V., Ngo, D., Todorov, K., "YAM++ Online: A Web Platform for Ontology and Thesaurus Matching and Mapping Validation," in The Semantic Web: ESWC 2017 Satellite Events, Portorož, Slovenia, 2017.
12. Jiménez-Ruiz, E., Cuenca Grau, B., "LogMap: Logic-Based and Scalable Ontology Matching," in The Semantic Web – ISWC 2011. Lecture Notes in Computer Science, Bonn, Germany, 2011.
13. Severo, B., Santos, C.T., Vieira, R., "VOAR: A Visual and Integrated Ontology Alignment Environment," in LREC, Reykjavik, Iceland, 2014.
14. Binding, C. and Tudhope. D., "Improving interoperability using vocabulary linked data," International Journal on Digital Libraries (IJDL), vol. 17, no. 1, pp. 5-21, 2016.

[13] http://researchspace.org/