# Identifying and Validating Ontology Mappings by Formal Concept Analysis

Mengyi Zhao[1] and Songmao Zhang[2]

[1,2]Institute of Mathematics, Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing, P. R. China
[1]myzhao@amss.ac.cn, [2]smzhang@math.ac.cn

**Abstract.** As a well developed mathematical model for analyzing individuals and structuring concepts, Formal Concept Analysis (FCA) has been applied to ontology matching (OM) tasks since the beginning of OM research, whereas ontological knowledge exploited in FCA-based methods is limited. The study in this paper aims to empowering FCA with as much as ontological knowledge as possible for identifying and validating mappings across ontologies. Our method, called FCA-Map, constructs three types of formal contexts and extracts mappings from the lattices derived. Firstly, the token-based formal context describes how class names, labels and synonyms share lexical tokens, leading to lexical mappings (anchors) across ontologies. Secondly, the relation-based formal context describes how classes are in taxonomic, partonomic and disjoint relationships with the anchors, leading to positive and negative structural evidence for validating the lexical matching. Lastly, after incoherence repair, the positive relation-based context can be used to discover additional structural mappings. Evaluation on anatomy track and large biomedical ontologies track of the 2015 Ontology Alignment Evaluation Initiative (OAEI) campaign demonstrates the effectiveness of FCA-Map and its competitiveness with 2015 OAEI top-ranked OM systems.
**Keywords:** ontology matching, Formal Concept Analysis, concept lattice.

## 1 Introduction

In the Semantic Web, ontologies model domain conceptualizations so that applications built upon them can interoperate with each other by sharing the same meanings. Such knowledge sharing and reuse can be severely hindered by the fact that ontologies for the same domain are often developed for various purposes, differing in coverage, granularity, naming, structure and many other aspects. Ontology matching (OM) techniques aim to alleviate the heterogeneity by identifying correspondences across ontologies. Ontology matching can be performed at the element level and the structure level [4]. The former considers ontology classes and their instances independently, such as string-based and language-based techniques, whereas the latter exploits relations among entities, including graph-based and taxonomy-based techniques. Most ontology matching systems [2,3,5,9,11] adopt both element and structure level techniques to achieve better performance.

Among the first batch of OM algorithms and tools proposed in the early 2000s, FCA-Merge [13] distinguished in using Formal Concept Analysis (FCA) formalism to

derive mappings from classes sharing textual documents as their individuals. Proposed by Rudolf Wille [14], FCA is a well developed mathematical model for analyzing individuals and structuring concepts. FCA starts with a formal context consisting of a set of objects, a set of attributes, and their binary relations. Concept lattice, or Galois lattice, can be computed based on formal context, where each node represents a formal concept composed of a subset of objects (extent) with their common attributes (intent). The extent and the intent of a formal concept uniquely determine each other in the lattice. Moreover, the lattice represents a concept hierarchy where one formal concept becomes sub-concept of the other if its objects are contained in the latter. FCA can naturally be applied to ontology construction [12], and is also widely used in data analysis, information retrieval, and knowledge discovery.

Following the steps of FCA-Merge, several OM systems continued to use FCA as well as its alternative formalisms, exploiting different entities as the sets of objects and attributes for constructing formal contexts [1, 8, 15]. FCA-OntMerge [8], for example, utilizes the classes of ontologies and their attributes to form its formal context, whereas in [1] the formal context is composed of ontology classes as objects and terms of a domain-specific thesaurus as attributes. Different types of formal contexts decide the information used for ontology matching, and we observed that some intrinsic and essential knowledge of ontology has not been involved yet, including both textual information within classes (e.g., class names, labels, and synonyms) and relationships among classes (e.g., ISA, sibling, and disjointedness relations).

This motivated the study in this paper, i.e., empowering FCA with as much as ontological information as possible for identifying and validating mappings across ontologies. Our method, called FCA-Map, generates three types of formal contexts and extracts mappings from the lattices derived. Firstly, the token-based formal context describes how class names, labels and synonyms share lexical tokens, leading to lexical mappings (anchors) across ontologies. Secondly, the relation-based formal context describes how classes are in taxonomic, partonomic and disjoint relationships with the anchors, leading to positive and negative structural evidence for validating the lexical matching. Lastly, after incoherence repair, the positive relation-based context can be used to discover additional structural mappings. Evaluation on anatomy track and large biomedical ontologies track of the 2015 Ontology Alignment Evaluation Initiative (OAEI) campaign demonstrates the effectiveness of FCA-Map and its competitiveness with 2015 OAEI top-ranked OM systems.

## 2 Preliminaries

Formal Concept Analysis (FCA) is a mathematical theory of data analysis using formal contexts and concept lattices. Formal context is defined as a triple $\mathbb{K} := (G, M, I)$, where $G$ is a set of objects, $M$ a set of attributes, and $I$ a binary relation between $G$ and $M$ in which $gIm$ holds, i.e., $(g, m) \in I$, reads: object $g$ has attribute $m$ [6]. Formal contexts are often illustrated in binary tables, as exemplified by Table 1, where rows correspond to objects, columns to attributes, and a cell is marked with "$\times$" if the object in its row has the attribute in its column.

**Definition 1.** *[6] For subsets of objects and attributes $A \in G$ and $B \in M$, derivation operators are defined as follows:*

$$A' = \{m \in M \mid gIm \; for \; all \; g \in A\}$$
$$B' = \{g \in G \mid gIm \; for \; all \; m \in B\}$$

$A'$ denotes the set of attributes common to the objects in $A$; $B'$ denotes the set of objects which have all the attributes in $B$.

A formal concept of context $\mathbb{K}$ is a pair $(A, B)$ consisting of extent $A \in G$ and intent $B \in M$ such that $A = B'$ and $B = A'$. $\mathfrak{B}(\mathbb{K})$ denotes the set of all formal concepts of context $\mathbb{K}$. The partial order relation, namely subconcept-superconcept-relation, is defined as:

$$(A_1, B_1) \leq (A_2, B_2) :\Leftrightarrow A_1 \subseteq A_2 (\Leftrightarrow B_1 \supseteq B_2)$$

Relation $\leq$ is called a hierarchical order of formal concepts. $\mathfrak{B}(\mathbb{K})$ ordered in this way is exactly a complete lattice, called the concept lattice and denoted by $\underline{\mathfrak{B}}(\mathbb{K})$.

| | vertebrate | mammal | flying | aquatic | carnivorous |
|---|---|---|---|---|---|
| elephant | × | × | | | |
| dolphin | × | × | | × | × |
| porpoise | × | × | | × | × |
| hawk | × | | × | | × |
| octopus | | | | × | × |

**Table 1:** An example formal context $\mathbb{K}_e$.

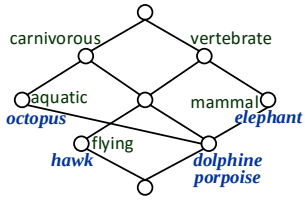


**Fig. 1:** Concept lattice $\underline{\mathfrak{B}}(\mathbb{K}_e)$ with simplified labelling.

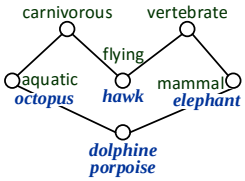


**Fig. 2:** GSH of concept lattice $\underline{\mathfrak{B}}(\mathbb{K}_e)$.

For an object $g \in G$, its *object concept* $\gamma g := (\{g\}'', \{g\}')$ is the smallest concept in $\underline{\mathfrak{B}}(\mathbb{K})$ whose extent contains $g$. In other words, object $g$ can generate formal concept $\gamma g$. Symmetrically, for an attribute $m \in M$, its *attribute concept* $\mu m := (\{m\}', \{m\}'')$ is the greatest concept in $\underline{\mathfrak{B}}(\mathbb{K})$ whose intent contains $m$. In other words, object $m$ can generate formal concept $\mu m$. For a formal concept $(A, B)$, its *simplified extent* (*simplified intent*), denoted by $K_{ex}$ ($K_{in}$), is a minimal description of the concept. Each object (attribute) in $K_{ex}$ ( $K_{in}$ ) can generate the formal concept $(A, B)$. As a matter of fact, $K_{ex}$ dose not appear in any descendant of $(A, B)$ and $K_{in}$ dose not appear in any ancestor of $(A, B)$. Figure 1 shows the concept lattice of context $\mathbb{K}_e$ in Table 1, where each formal concept is labeled by its simplified extent and intent.

Galois Sub-hierarchy (GSH) introduced by [7] is a sub-structure of concept lattice. Only concepts carrying information are retained in GSH, meaning that GSH solely contains formal concepts that introduce new objects or new attributes and excludes formal concepts whose $K_{ex}$ and $K_{in}$ are both empty. The ordering of formal concepts in GSH is the same as in the original concept lattice. Removing the formal concepts without labels in Figure 1 leads to the GSH shown in Figure 2.

## 3 The FCA-Map Method

Given two ontologies, FCA-Map builds formal contexts and uses the derived concept lattices to cluster the commonalities among ontology classes, at lexical level and structural level, respectively. Concretely, FCA-Map performs step-by-step as follows.

1. **Acquiring anchors lexically.** The token-based formal context is constructed, and from its derived concept lattice, a group of lexical anchors $\mathcal{A}$ across ontologies can be extracted.
2. **Validating anchors structurally.** Based on $\mathcal{A}$ , the relation-based formal context is constructed, and from its derived concept lattice, positive and negative structural evidence of anchors can be extracted. Moreover, an enhanced alignment $\mathcal{A}'$ without incoherences among anchors is obtained.
3. **Discovering additional matches.** Based on $\mathcal{A}'$, the positive relation-based formal context is constructed, and from its derived concept lattice, additional matches across ontologies can be identified.

We take two anatomical ontologies, Adult Mouse Anatomy[1] (MA) and the anatomy subset of National Cancer Institute Thesaurus[2] (NCI), to demonstrate our method. MA is a structured controlled vocabulary describing the anatomical structure of the adult mouse, whereas NCI describes the human anatomy for the purpose of cancer research. The versions used are the OWL files of these two ontologies provided by the 2015 OAEI. For MA and NCI, the token-based and relation-based formal contexts are of large-size, resulting in complex structures of the concept lattices derived. In order to avoid generating redundant information, GSH, a polynomial-sized representation of concept lattice that preserves the most pertinent information, is utilized in FCA-Map.

### 3.1 Constructing the token-based formal context to acquire lexical anchors

Most OM systems rely on lexical matching as initiation due to the fact that classes sharing names across ontologies quite likely represent the same entity in the domain of interest. FCA-Map, rather than using lexical and linguistic analysis, generates a formal context at the lexical level and obtains mappings from the lattice derived from the context.

The token-based formal context $\mathbb{K}_{lex} := (G_{lex}, M_{lex}, I_{lex})$ is described as follows. Names of ontology classes as well as their labels and synonyms, when available, are exploited after normalization that includes inflection, tokenization, stop word elimination[3], and punctuation elimination. In $\mathbb{K}_{lex}$, $G_{lex}$ is the set of strings each corresponding to a name, label, or synonym of classes in two ontologies, $M_{lex}$ is the set of tokens in these strings, and binary relation $(g, m) \in I_{lex}$ holds when string $g$ contains token $m$,

[1] http://www.informatics.jax.org/glossary/adult_ma_dictionary

[2] https://ncit.nci.nih.gov/ncitbrowser/

[3] Although eliminating the stop words carrying logical meanings may affect the precision, its benefit in recall is more advantageous according to our experiments.

or a synonym[4] or lexical variation[5] of $m$. Table 2 shows $\mathbb{K}_{lex}$ of a small part of MA and NCI, and its derived concept lattice in GSH form is displayed in Figure 3. For each formal concept derived, in addition to strings in its extent, we are also interested in the classes that these strings come from, called *class-origin extent*. For example, in Figure 3, the *class-origin extent* of formal concept by node 7 is {MA:*mammary gland fluid/secretion*, NCI:*Breast Fluid or Secretion*} since in NCI, "*Mammary Gland Fluids and Secretions*" is a synonym of class NCI:*Breast Fluid or Secretion*.

| | gland | adrenal | zona | zone | fasciculata | reticularis | salivary | palatine | mammary | secretion | fluid |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MA:*palatine gland* | × | | | | | | | × | | | |
| MA:*adrenal gland zona fasciculata* | × | × | × | | × | | | | | | |
| MA:*adrenal gland zona reticularis* | × | × | × | | | × | | | | | |
| MA:*mammary gland fluid/secretion* | × | | | | | | | | × | × | × |
| NCI:*Palatine Salivary Gland* | × | | | | | | × | × | | | |
| NCI:*Fasciculata Zone* | | | | × | × | | | | | | |
| NCI:*Reticularis Zone* | | | | × | | × | | | | | |
| NCI:*Mammary Gland Fluids and Secretions* | × | | | | | | | | × | × | × |

**Table 2:** Token-based formal context $\mathbb{K}_{lex}$ of a small part of MA and NCI.

An essential property of FCA is the duality between a set of objects and their attributes. The more attributes demanded, the fewer objects can meet the requirements. In the case of the token-based formal concept, the more common tokens appearing in its intent, the fewer strings the extent contains, and the more possibly for the classes in *class-origin extent* to be matched. This is to say that cardinality of the extent can reflect how similar the strings are, thus classes from different source ontologies in a smaller-sized *class-origin extent* can be considered as a mapping with higher confidence. Practically, we restrict our attention to formal concepts whose *simplified extent* or *class-origin extent* contains exactly two strings or classes across ontologies, and extract two types of lexical anchors, namely ***Type I anchor*** for the exact match, and ***Type II anchor*** for the partial match, respectively. Of note, on the other hand, cardinality of the intent cannot be used to measure the similarity of strings. For example, MA:*nerve* and NCI:*Nerve*, which is a match, only share one token, whereas MA:*left lung respiratory bronchiole* and NCI:*Right Lung Respiratory Bronchiole*, not a match, share three tokens.

***Type I anchor***. *Simplified extent* $K_{ex}$ of the formal concept contains exactly two strings from classes across ontologies. This indicates that the two strings are composed of the same or synonymous tokens, thus the corresponding classes are extracted to be a match, as exemplified by $(MA$ : *mammary gland fluid/secretion*, $NCI$ : *Breast Fluid or Secretion*$)$ through formal concept of node 7 in Figure 3 whose $K_{ex}$ has two strings, one from MA and the other NCI.

***Type II anchor.*** The *class-origin extent* of the formal concept contains exactly two classes across ontologies and *simplified extent* $K_{ex}$ contains strings from at most

[4] Sub-Term Mapping Tools (https://lsg2.nlm.nih.gov/LexSysGroup/Projects/stmt/2013+/web/index.html) are used to access synonyms.

[5] SPECIALIST Lexicon (https://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/web/index.html) of UMLS is used to access lexical variations.

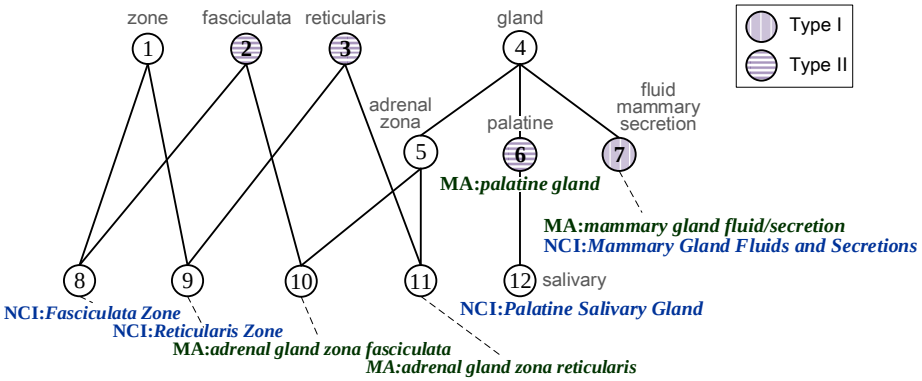


**Fig. 3:** Concept lattice in GSH with simplified labeling derived from $\mathbb{K}_{lex}$ in Table 2.

one source ontology. Here the strings share tokens in the intent rather than composed of the same or synonymous tokens. For example, (MA:*adrenal gland zona fasciculata*, NCI:*Fasciculata Zone*) is extracted from node 2 in Figure 3, due to the common token "fasciculata" which exists solely in these two classes. And (MA:*palatine gland*, NCI:*Palatine Salivary Gland*) is identified as an anchor from node 6, due to the common tokens "palatine" and "gland" which co-exist solely in these two classes.

### 3.2 Constructing the relation-based formal context to validate lexical anchors

Structural relationships of ontologies are exploited to validate the matches obtained at the lexical level. One of our previous studies [16] proposed using positive and negative structural evidence among anchors for the purpose of validation. More precisely, classes of one anchor sharing relationships to classes in another anchor can be seen as their respective positive evidence. On the other hand, negative structural evidence refers to the conflict based on the disjointedness relationships between classes. In FCA-Map, we build the relation-based formal context to obtain both positive and negative structural evidence for lexical anchors. Both explicitly represented and inferred semantic relations are used in our method.

| | (ISA) (MA:*ligament*, NCI:*Ligament*) | (I-D) (MA:*organ system*, NCI:*Organ System*) | (SIB) (MA:*adipose tissue*, NCI:*Adipose Tissue*) | (SIB) (MA:*larynx ligament*, NCI:*Laryngeal Ligament*) | (PAT) (MA:*larynx*, NCI:*Larynx*) |
|---|---|---|---|---|---|
| MA:*ligament* | | × | × | | |
| MA:*periodontal ligament* | × | × | | × | |
| MA:*auricular ligament* | × | × | | × | |
| MA:*adipose tissue* | | × | | | |
| MA:*larynx ligament* | × | × | | | × |
| NCI:*Ligament* | | × | | | |
| NCI:*Periodontium* | × | × | | × | |
| NCI:*Broad Ligament* | × | × | | × | |
| NCI:*Adipose Tissue* | | × | | | |
| NCI:*Laryngeal Ligament* | × | × | | | × |

**Table 3:** Relation-based formal context $\mathbb{K}_{rel}$ of a small part of MA and NCI.

The relation-based formal context $\mathbb{K}_{rel} := (G_{rel}, M_{rel}, I_{rel})$ is described as follows. Classes in two source ontologies are taken as object set $G_{rel}$, and lexical anchors prefixed with different relational labels are taken as attribute set $M_{rel}$. In the case of MA and NCI, four kinds of relationships are considered, *ISA*, *SIBLING-WITH*, *PART-OF*, and *DISJOINT-WITH*, labeled by "(ISA)", "(SIB)", "(PAT)", and "(I-D)" (or "(D-I)"), respectively. Binary relation $(g, m) \in I_{rel}$ holds if $g$ has the corresponding relationship (as in the prefix of $m$) with the class from the same source ontology as $g$ in the anchor of $m$. The relation-based formal context $\mathbb{K}_{rel}$ of a small part of MA and NCI is displayed in Table 3. For instance, MA:*periodontal ligament* and NCI:*Periodontium* are subclasses of MA:*ligament* and NCI:*Ligament*, respectively, thus (MA:*periodontal ligament*, (ISA)(MA:*ligament*, NCI:*Ligament*)) $\in I_{rel}$ and (NCI:*Periodontium*, (ISA)(MA: *ligament*, NCI:*Ligament*)) $\in I_{rel}$ hold. Moreover, MA:*adipose tissue* is a subclass of MA:*organ system* whereas NCI:*Adipose Tissue* is disjoint with NCI:*Organ System*, thus (MA:*adipose tissue*, (I-D)(MA:*organ system*, NCI:*Organ system*)) $\in I_{rel}$ and (NCI:*Adipose Tissue*, (I-D)(MA:*organ system*, NCI:*Organ system*)) $\in I_{rel}$ hold.

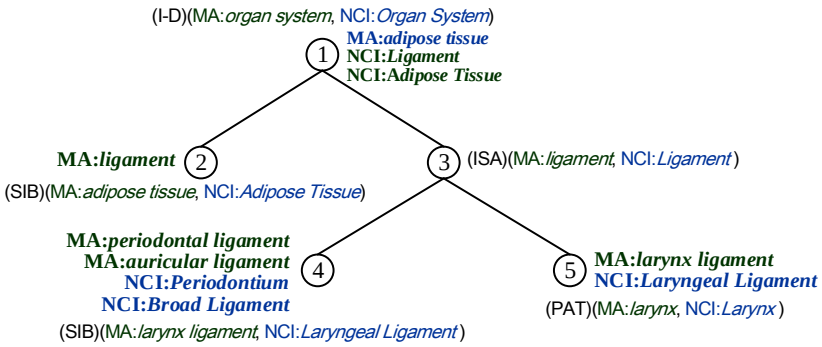


**Fig. 4:** GSH of $\mathbb{K}_{rel}$ with simplified labeling.

The derived concept lattice in GSH form of $\mathbb{K}_{rel}$ of a small part of MA and NCI is illustrated in Figure 4. Formal concepts whose extents include both classes in some anchors indicate structural evidence. Such anchors are positive evidence to anchors with label"(ISA)", "(SIB)" or "(PAT)" in the intent, and vice versa. Conversely, they are negative evidence to anchors with label "(I-D)" or "(D-I)" in the intent, and vice versa. In this way, positive and negative structural evidence set of each anchor $a$ can be obtained, denoted by $P(a)$ and $N(a)$, respectively. For example, in the extent of node 3 in Figure 4, (MA:*periodontal ligament*, NCI:*Periodontium*) and (MA:*larynx ligament*, NCI:*Laryngeal Ligament*), two anchors acquired lexically, are positive evidences to anchor (MA:*ligament*, NCI:*Ligament*) with label "(ISA)" in the intent, and negative evidences to anchor (MA:*organ system*, NCI:*Organ System*) with label "(I-D)". The *support degree* and *incoherence degree* of each anchor are the cardinality of its positive and negative evidence set, respectively.

Now we can utilize all the positive evidence sets $\mathcal{P}$ and negative evidence sets $\mathcal{N}$ to eliminate incorrect lexical anchors and retain the correct ones. There are two steps conducted one-by-one as follows.

*Incoherence repairing.* The negative evidence leads to incoherency among anchors, for which FCA-Map repairs in a greedy way, i.e., eliminating the incoherence-causing

anchors iteratively until $\mathcal{N}$ becomes empty. At each iteration, anchor $a$ having the least negative evidence set, i.e., the smallest *incoherence degree*, is selected. For every anchor $a'$ in $N(a)$, if *incoherence degree* of $a'$ is greater than $a$, eliminate $a'$; otherwise, compare the *support degree* of $a$ and $a'$, and eliminate the one with smaller *support degree*.

*Anchor screening.* Anchors having no positive structural evidence according to the updated $\mathcal{P}$ are either caused by the structural isolatedness of classes, or simply incorrect mismatches. FCA-Map screens anchors based on both lexical and structural evidence, where ***Type II anchors*** without positive evidence are eliminated.

### 3.3 Constructing the positive relation-based formal context to discover additional matches

After incoherence repair and screening, anchors retained are those supported both lexically and structurally. Based on the enhanced alignment, FCA-Map goes further to build the positive relation-based formal context aiming to identify new, structural mappings. The way positive relation-based formal context $\mathbb{K}'_{rel}$ constructed is similar to $\mathbb{K}_{rel}$, i,e., using classes in two source ontologies as object set and anchors prefixed with relationship labels as attribute set. In the case of MA and NCI, five kinds of relationships are considered, *ISA*, *SUPERCLASS-OF*, *SIBLING-WITH*, *PART-OF*, and *HAS-PART*, where disjointedness relationship is no longer necessary. For the derived formal concepts, we restrict our attention to those with exactly two classes across ontologies in the *simplified extent*. Although most of the mappings extracted this way have already been identified at the lexical level, new additional matches emerge, as exemplified by $($MA: *hindlimb bone*, NCI: *Bone of the Lower Extremity*$)$.

## 4 Evaluation

To demonstrate the effectiveness of FCA-Map, evaluation is performed on two pairs of real-world ontologies, Adult Mouse Anatomy (2,744 classes) and the anatomy subset of NCI Thesaurus (3,304 classes); and the Foundational Model of Anatomy (3,696 classes) and NCI (6,488 classes), respectively, from anatomy track and large biomedical ontologies track of OAEI 2015. FCAlib[6] is used to derive concept lattices (GSH) from formal contexts. It is an open-source, extensible library for FCA tool developers. FCA-Map is implemented in Java and the experiments were conducted in a PC with Intel i7 (3.60GHz) and 8GB RAM. It took 166 seconds and 425 seconds, respectively, for FCA-Map to finish the MA-NCI$_{Anat.}$ and the FMA-NCI matching.

### 4.1 Anchors obtained

The results of lexical matching by FCA-Map are summarized in Table 4, and structural matching is presented in Table 5 where the upper part is about structural validation and the lower part about extra discovered structural mappings. Columns "Corr.", "Incor.", and "Unkn." indicate the number of correct, incorrect, and unknown mappings, respectively, as categorized by OAEI where "unknown" mappings will neither be considered as correct nor incorrect when evaluating the alignment, but will simply be ignored.

| | MA-NCI$_{Anat.}$ | | | | FMA-NCI | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Types of anchors | **Total** | Corr. | Incor. | P | **Total** | Corr. | Unkn. | Incor. | P |
| *Type I* | 1,223 | 1,163 | 60 | 95.1% | 2,759 | 2,416 | 248 | 95 | 96.2% |
| *Type II* | 172 | 113 | 59 | 65.7% | 131 | 60 | 4 | 67 | 47.2% |
| **Total** | 1,395 | 1,276 | 119 | 91.5% | 2,890 | 2,476 | 252 | 162 | 93.9% |

**Table 4:** Results of lexical anchors.

| | MA-NCI$_{Anat.}$ | | | | FMA-NCI | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Types of anchors | **Total** | Corr. | Incor. | P | **Total** | Corr. | Unkn. | Incor. | P |
| *Type I* | 1,220 | 1,161 | 59 | 95.2% | 2,703 | 2,414 | 208 | 81 | 96.8% |
| *Type II* | 125 | 98 | 27 | 78.4% | 63 | 46 | 2 | 15 | 75.4% |
| **Total** | 1,345 | 1,259 | 86 | 93.6% | 2,766 | 2,460 | 210 | 96 | 96.2% |
| Additional | 16 | 10 | 6 | 62.5% | 25 | 3 | 0 | 22 | 12% |
| **Total** | 1,361 | 1,269 | 92 | 93.2% | 2,791 | 2,463 | 210 | 118 | 95.4% |

**Table 5:** Results of enhanced alignment.

One can see that most of the lexical anchors are of ***Type I***, i.e., the name, synonym or label of one class is the same as another class. For example, MA:*cortical layer II* and NCI:*External Granular Layer* are extracted as an anchor because in MA, "*external granular layer*" is a synonym of MA:*cortical layer II*. Incorrect ***Type I anchors*** mainly come from three cases. (1) Although having the same name, classes in anchor do not represent equivalent entity. For example, MA:*organ system* and NCI:*Organ System*, although sharing matched subclasses, have respective additional different subclasses. (2) Mismatched classes may be considered to be a mapping based on their synonyms or labels. For example, anchor (MA:*cerebellum lobule I*, NCI:*Lingula* ) (through synonym "*lingula*" in MA) is a mismatch because the former is a part of cerebellar vermis and the latter a part of left lung. (3) Using external lexicon may introduce incorrect anchors. For example, MA:*back* matches NCI:*Dorsum* because "back" and "dorsum" are synonymous according to the lexicon used in FCA-Map. This is a mismatch because in MA back is a part of trunk, while in NCI dorsum refers to outer surface of scapula.

***Type II*** lexical anchors have lower precisions, reflecting the unstable performance of relying on names sharing tokens to derive commonalities of classes. Nevertheless, many incorrect anchors can be eliminated in the validation process, causing the precision to increase, for instance from 47.2% to 75.4% for ***Type II*** anchors in FMA-NCI. Take ***Type II*** anchor (MA:*retina ganglion cell layer*, NCI: *Retinal Ganglion Cell*) for example. It is eliminated in incoherence repair because of its conflict with (MA:retina layer, NCI: Retina Layer), of which the *support degree* is 0 and 8, respectively. The structural validation based on the relation-based concept lattice in FCA-Map can ensure to improve the precision of lexical mappings.

### 4.2 Comparing with other lexical matching methods

Among many lexical matching methods such as string equality, substring test, and edit distance, TFIDF-based methods [4] are of particular interest because similarly to FCA-Map they are based on tokens. Adopted in OM systems YAM++ [3] and GMap [10],

[6] https://julianmendez.github.io/fcalib/

TFIDF measures simultaneously how often the tokens appear in one class name and how much information the tokens bring across names of classes from different ontologies. We compare the performance of lexical matching of FCA-Map with TFIDF solely using the class names of MA and NCI without any external resources. The result is shown in Figure 5, where F-measure of FCA-Map is higher than TFIDF for any threshold.

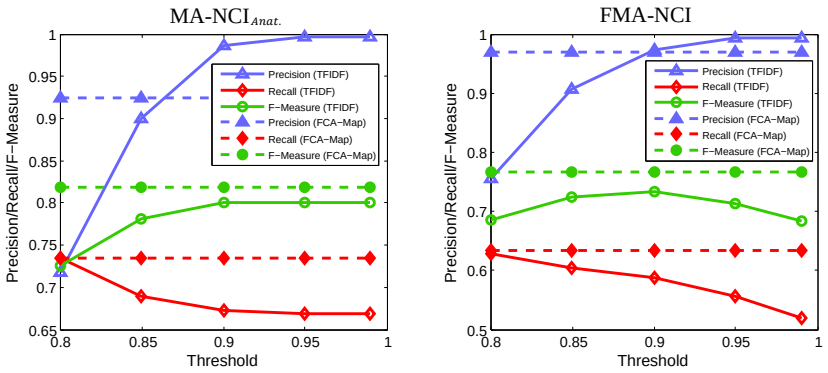


**Fig. 5:** Comparing with TFIDF.

Compared with the TFIDF-based methods, FCA-Map emphasizes on the particular commonality of two strings, and there is no need for setting thresholds which is required in TFIDF for selecting matches. This can be illustrated by MA: *tectum* and NCI: *tectum mesencephali*. They are not matched according to TFIDF because token "mesencephali" has a high inverse-document-frequency (it solely appears in this string) and token "tectum" is ignored (it solely appears in the two strings). On the other hand, this correspondence can be derived in our method since there is a formal concept with intent {"tectum"} and extent exactly containing these two strings. Moreover, our method can avoid the mistake of locally measuring frequency of tokens. For instance, MA: *common iliac artery* and NCI: *Right Common Iliac Artery* have a relatively high similarity (0.86) according to TFIDF, while this pair is not extracted by FCA-Map. There are many other class names share tokens "common", "iliac", and "artery", such as MA: *Left Common Iliac Artery* and NCI: *Right Common Iliac Artery Branch*, therefore what the two strings in comparison share are not unique enough for them to be chosen as a match. Indeed, our method features in detecting the particular commonality solely belongs to the names compared while ignoring the commonality shared by many other names.

### 4.3 Comparing with OAEI 2015 top-ranked systems

A comparison between FCA-Map and OAEI 2015 top-ranked systems is shown in Table 6. For MA-NCI$_{Anat.}$, the precision, recall and F-measure of FCA-Map ranks second, fifth, and forth, respectively. Results of FMA-NCI are encouraging, with both recall and F-measure tie for first. Moreover, FCA-Map is capable of extracting mappings that cannot be identified by other systems, as exemplified by ***Type II*** anchors (MA:*adrenal gland zona reticularis*, NCI:*Reticularis Zone*), (MA:*ileocaecal junction*, NCI:*Ileocecal Valve*). These mappings are identified in the token-based concept lattice and validated

in the relation-based concept lattice. The tokens shared by two classes in these mappings are unique to their names. The lexical matching method of FCA-Map is suitable for domain ontologies having class names, labels, or synonyms from domain-specific vocabulary, whereas its performance can be relatively poor for general-purpose ontologies whose terminologies are more varied and ambiguous, like those in the conference track of OAEI where FCA-Map ranked at the average level. Additionally, for negative evidence to be identified, our method requires that at least one source ontology declares disjointedness relationships between classes.

| | MA-NCI$_{Anat.}$ | | | FMA-NCI | | |
|---|---|---|---|---|---|---|
| Systems | P | R | F | P | R | F |
| XMAP-BK | - | - | - | 0.971 | 0.902 | 0.935 |
| AML | 0.956 | 0.931 | 0.944 | 0.960 | 0.899 | 0.928 |
| LogMap | 0.918 | 0.846 | 0.88 | 0.949 | 0.901 | 0.924 |
| LogMapBio | 0.882 | 0.901 | 0.891 | 0.926 | 0.917 | 0.921 |
| XMAP | 0.928 | 0.865 | 0.896 | 0.970 | 0.784 | 0.867 |
| FCA-Map | 0.932 | 0.837 | 0.882 | 0.954 | 0.917 | 0.935 |

**Table 6:** Comparing with OAEI 2015 top-ranked systems.

## 5 Discussion and Conclusions

**Discovering complex mappings structurally.** As shown in Table 5, structural mappings identified by the positive relation-based concept lattice are limited. Nevertheless, in the lattice we noticed that the *simplified extents* of some formal concepts contain more than two classes from different source ontologies, meaning these classes share the same structural relationships to anchors in the intent. Such classes may compose a complex mapping, as elaborated in the following.

1. *One-to-group mappings*. The *simplified extent* contains only one class from one source ontology and multiple classes from the other source ontology. For example, MA:*inferior suprarenal vein* can be mapped to the group of concepts {NCI:*Left Suprarenal Vein*, NCI:*Right Suprarenal Vein*} as the three concepts are contained within one *simplified extent* that has no more classes. This one-to-group mapping comes from the difference in granularity between MA and NCI.
2. *Group-to-group mappings*. The *simplified extent* contains multiple classes from different source ontologies, respectively. For example, two groups of concepts {MA:*sacral vertebra 1*, MA:*sacral vertebra 2*, MA:*sacral vertebra 3*, MA:*sacral vertebra 4*} and {NCI:*S1 Vertebra*, NCI:*S2 Vertebra*, NCI:*S3 Vertebra*, NCI:*S4 Vertebra*, NCI:*S5 Vertebra*} can be mapped as these classes are contained in one *simplified extent* that has no more classes. This group-to-group mapping represents the difference between mouse and human anatomy.

Compared with other FCA-based OM systems, the study in this paper is more comprehensive as an attempt to push the envelope of the Formal Concept Analysis formalism in ontology matching tasks. Three types of formal contexts are constructed one-by-one, and their derived concept lattices are used to cluster the commonalities among

classes at lexical and structural level, respectively. Experiments on large, real-world domain ontologies show promising results and reveal the power of FCA. Our future work would introduce more elements of ontology into FCA-Map including properties, individuals, and logical constructors and axioms. Optimization techniques for handling large-scale FCA contexts will also be worth exploring.

**Acknowledgements.** This work has been supported by the National Key Research and Development Program of China under grant 2016YFB1000902, the Natural Science Foundation of China under No. 61232015, the Knowledge Innovation Program of the Chinese Academy of Sciences (CAS), Key Lab of Management, Decision and Information Systems of CAS, and Institute of Computing Technology of CAS.

## References

1. de Souza, K.X.S., Davis, J.: Aligning ontologies and evaluating concept similarities. In: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", Springer (2004) 1012–1029
2. Djeddi, W.E., Khadir, M.T.: Xmap: a novel structural approach for alignment of owl-full ontologies. In: Machine and Web Intelligence (ICMWI), 2010 International Conference on, IEEE (2010) 368–373
3. Duyhoa, N., Bellahsene, Z.: Yam++ results for oaei 2012. In: Seventh International Workshop on Ontology Matching. (2012) 226–233
4. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer Science & Business Media (2013)
5. Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I.F., Couto, F.M.: The agreementmakerlight ontology matching system. In: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", Springer (2013) 527–541
6. Ganter, B., Wille, R.: Formal concept analysis: mathematical foundations. Springer Science & Business Media (2012)
7. Godin, R., Mili, H.: Building and maintaining analysis-level class hierarchies using galois lattices. In: ACM SIGplan Notices. Volume 28., ACM (1993) 394–410
8. Guan-yu, L., Shu-peng, L., et al.: Formal concept analysis based ontology merging method. In: Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on. Volume 8., IEEE (2010) 279–282
9. Jiménez-Ruiz, E., Grau, B.C.: Logmap: Logic-based and scalable ontology matching. In: International Semantic Web Conference, Springer (2011) 273–288
10. Li, W.: Combining sum-product network and noisy-or model for ontology matching. Ontology Matching (2015) 35
11. Niepert, M., Meilicke, C., Stuckenschmidt, H.: A probabilistic-logical framework for ontology matching. In: AAAI, Citeseer (2010)
12. Obitko, M., Snsel, V., Smid, J.: Ontology design with formal concept analysis. CLA **128**(3) (2004) 1377–1390
13. Stumme, G., Maedche, A.: Fca-merge: Bottom-up merging of ontologies. In: IJCAI. Volume 1. (2001) 225–230
14. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: Ordered sets. Springer (1982) 445–470
15. Xu, X., Wu, Y., Chen, J.: Fuzzy fca based ontology mapping. In: 2010 First International Conference on Networking and Distributed Computing, IEEE (2010) 181–185
16. Zhang, S., Bodenreider, O.: Experience in aligning anatomical ontologies. International journal on Semantic Web and information systems **3**(2) (2007) 1