Fraunhofer FOKUS

FRAUNHOFER INSTITUTE FOR OPEN COMMUNICATION SYSTEMS FOKUS

STUDY – CORPORATE SMART CONTENT EVALUATION

Technical Report TR-B-16-02

STUDY – CORPORATE SMART CONTENT EVALUATION

Technical Report TR-B-16-02

Ralph Schäfermeier (FU Berlin)
Alexandru-Aurelian Todor (FU Berlin)
Alexandra La Fleur (FU Berlin)
Ahmad Hasan (FU Berlin)
Johannes Einhaus (Fraunhofer FOKUS)
Adrian Paschke (Fraunhofer FOKUS)

Berlin, 31.05.2016

Contents

Contents

1 Management Summary

Nowadays, a wide range of information sources are available due to the evolution of web and collection of data. Plenty of these information are consumable and usable by humans but not understandable and processable by machines. Some data may be directly accessible in web pages or via data feeds, but most of the meaningful existing data is hidden within deep web databases and enterprise information systems. Besides the inability to access a wide range of data, manual processing by humans is effortful, error-prone and not contemporary any more. Semantic web technologies deliver capabilities for machine-readable, exchangeable content and metadata for automatic processing of content. The enrichment of heterogeneous data with background knowledge described in ontologies induces re-usability and supports automatic processing of data.

The establishment of *"Corporate Smart Content"* (CSC) - semantically enriched data with high information content with sufficient benefits in economic areas - is the main focus of this study. We describe three actual research areas in the field of CSC concerning scenarios and datasets applicable for corporate applications, algorithms and research.

Aspect-oriented Ontology Development advances modular ontology development and partial reuse of existing ontological knowledge. **Complex Entity Recognition** enhances traditional entity recognition techniques to recognize clusters of related textual information about entities. **Semantic Pattern Mining** combines semantic web technologies with pattern learning to mine for complex models by attaching background knowledge.

This study introduces the afore-mentioned topics by analyzing applicable scenarios with economic and industrial focus, as well as research emphasis. Furthermore, a collection of existing datasets for the given areas of interest is presented and evaluated. The target audience includes researchers and developers of CSC technologies - people interested in semantic web features, ontology development, automation, extracting and mining valuable information in corporate environments.

The aim of this study is to provide a comprehensive and broad overview over the three topics, give assistance for decision making in interesting scenarios and choosing practical datasets for evaluating custom problem statements. Detailed descriptions about attributes and metadata of the datasets should serve as starting point for individual ideas and approaches.

1 Management Summary

2
Test Scenarios

This section introduces the three topics of Aspect-oriented Knowledge Engineering, Complex Entity Recognition and Semantic Pattern Mining and describes test scenarios concerning industry and research.

2.1
Aspect-oriented Knowledge Engineering

2.1.1
Overview

With *Aspect-Oriented Ontology Development* we refer to a methodological approach, a set of logical formalisms and accompanying tools for modular ontology development and partial reuse of existing ontological knowledge based on requirements and cross-cutting concerns[1].

It is therefore a subfield of *Ontology Engineering* and a particular approach to the problem of *Ontology Modularization* and *Modular Ontology Development*.

Ontologies are an effective means for normalizing concepts and relations between concepts. Ontologies may be used to formally and explicitly describe assumptions about a domain of interest made by a particular agent. An ontology about a particular domain may in turn be shared with the public (i.e. other agents) who may use the shared knowledge explicated by them. Ontology developers may extend or alter existing ontologies.

Ontology development constitutes a process which in most of the cases involves more than one group of protagonists. Normally these are ontology experts who have experience in using formal knowledge representation languages, tools, and methods as well as domain experts possessing knowledge of the facts concerning the domain of interest.

While parts of the world's knowledge are static (ground truth), other parts may be subject to evolutionary change. Methods for ontology development must take this dynamic nature of knowledge into consideration.

Ontology engineering refers to a set of activities concerning the process of creation and lifecycle of ontologies. In detail, it refers to methodologies for the creation of ontologies and complementing tools and languages.

2.1.2
Aspect-Oriented Ontology Development

The main goal of Aspect-Oriented Programming is the decomposition of software systems into concerns which cross-cut the system. A code module covering a particular concern is referred to as an aspect. Concerns may be functional concerns, which are directly related to the systems's domain of interest and business logic and non-functional concerns, such as security, logging/auditing and performance.

The decomposition is accomplished by introducing extensions to existing programming languages (such as AspectJ[2] for Java) that allow the decomposition of code into modules, each of them dealing with a concern, as well as a mechanism for recombining the modules at compile or runtime into a complete and coherent system. Programming languages without

1 As defined by the IEEE standard 1471 of software architecture [Gro00], *"concerns are those interests which pertain to the system's development, its operation or any other aspects that are critical or otherwise important to one or more stakeholders".*

2 `https://eclipse.org/aspectj/`

aspect-orientation have no means for separating those concerns, which leads to undesired code tangling and hinders system decomposition.

Two principle properties of Aspect-Oriented Programming are *quantification* and *obliviousness* [FF00]. *Obliviousness* refers to the fact that all information necessary to determine the execution points where the application should make a call into an aspect module are contained within the aspect itself rather than in the application code. A developer of one module does not, and need not, have knowledge about other modules that might potentially be called.

This information may be provided in the form of an exhaustive list of signatures or in terms of *quantified statements* over signatures, called a *pointcut*. Each single matching signature is called a *join point*.

Formally, Aspect-Oriented Programming uses quantified statements of the following form [Ste05]:

$$\forall m(p_1, \ldots, p_n) \in M : s(m(p_1, \ldots, p_n)) \qquad \rightarrow (m(p_1, \ldots, p_n) \rightarrow a(p_1, \ldots, p_n)), \qquad \text{Formula 2.1}$$

where M is the set of all methods defined in the software system, s a predicate specifying a matching criterion, $m(p_1, \ldots, p_n) \in M$ a method matching the signature $m(p_1, \ldots, p_n)$, and $a(p_1, \ldots, p_n)$ the execution of the aspect with all the parameters of each method, respectively. The code in the aspect, which is executed at each joint point, is referred to as *advice*. In APO terminology, an aspect *advices* the main code.

The idea behind Aspect-Oriented Ontology Development is to use pointcuts in order to describe ontology modules and aspects in order to attach additional knowledge (advice) to each of these modules.

The semantics of ontology aspects are defined in correspondence with the possible-world semantics of multi-modal logics:

- Aspects correspond to sets of axioms or facts that are true in certain possible worlds.
- Aspects are modeled as classes.
- Possible worlds are modeled as individuals.
- Accessibility relations are modeled as object properties.
- The semantics of aspects depend on the choice of conditions on frames (axioms on accessibility properties).

The rationales behind that choice are explained as follows:

- (Multi-)modal logics are a syntactic variant of and thereby semantically equivalent to Description Logics [Sch91, BCM$^+$03].
- Aspects are a sort of modality in that there is a function that determines in which situations an aspect is active and in which it is not. That corresponds to possible worlds in modal logics where a truth-functional valuation determines whether a fact is valid in a possible world or not.
- The kind of modal logic is determined by conditions on Kripke frames, which (to a certain extent) may be controlled by fixing the characteristics of the accessibility relations. This allows the representation of e.g., temporal logic (as in our running example), simple views, agent beliefs, etc.
- Using classes as aspects allows to use abstract class definitions using constraints with quantifiers.

Defining ontology aspects as meta-statements which describe possible worlds in which an ontology axiom may be either true or false allows for a number of industrial as well as research scenarios which are described in the following section.

2.1.3 Research Scenarios

- Simple views
  Aspects in their simplest form provide simple context information with attach to ontology axioms or facts.
  Simple views are realized using a simple Logic K with only one condition on the modal frames:

$$(M) : \Box A \rightarrow A \qquad \text{Formula 2.2}$$

  This modal axiom is equivalent to having a *reflexive* accessibility relation in the Kripke frame. Figure 1 shows the general structure of a realization of simple views using reflexive relations.

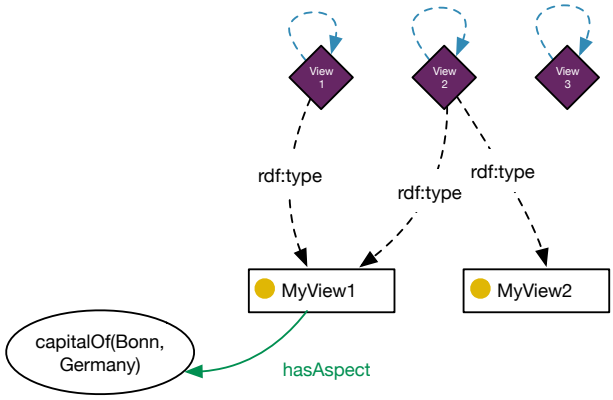

Figure 1: Simple non-contradicting views realized by using the semantics of a modal logic K.

  With Logic K it is possible to express:
  - named views
  - provenance information
  - meta-data of any type
  - references to other named resources, e.g. for
    - multi-faceted alignment to external taxonomies
    - optional, non-normative descriptions

  The standard way of solving multi-faceted classification is be following the view inheritance ontology design pattern[3]. One shortcoming of this pattern is, however, that the classifiers become part of the inheritence hierarchy and thereby of the domain conceptualiztion itself, while in fact, they should belong to the meta-level.
  Figure 2 shows how aspects may be used to have multiple classification hiearchies with external classifiers as ontology aspects, as well as how it is accomplished using the view inheritance pattern.
- Inconsistent knowledge
  Logic K is an epistemic logic and relies on the fact that knowledge is consistent across the boundaries of possible worlds. This scenario also describes cases in which the axiomatization of a domain may change depending on an externally provided context, but, unlike in the first scenario, inconsistencies in the knowledge in different contexts is allowed.
  An example of a inconsistency would be two contradicting axioms in the same ontology, each of which represents the view of a particular stakeholder. Figure 3 shows an example of the axiomatic description of the concept *vehicle* as seen by two different agents or stakeholders. The two axioms Vehicle $\equiv$ hasWheel exactly 2 and Vehicle $\equiv$ hasWheel min 4 contradict each other, and introducing both of them into

3 `http://ontologydesignpatterns.org/wiki/Submissions:View_Inheritance`

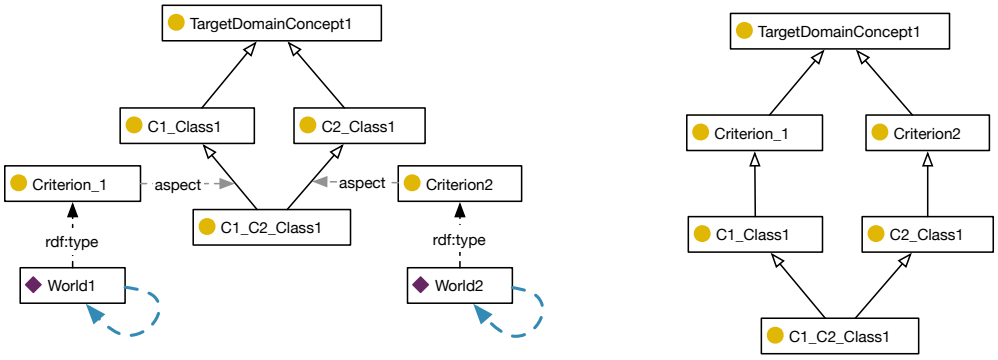

Figure 2: Multiple-view inheritance using external classifiers with aspects (left) and following the view inheritance pattern (right).

the same knowledge base would render the entire knowledge base inconsistent, because of the principle of explosion *ex contradictione sequitur quodlibet*.
A possible way of reasoning with inconsistent knowledge is resorting to paraconsistent logic and consists in preventing the explosion at the cost of abandoning one of the three principles disjunction induction ($A \vdash A \vee B$), the disjunctive syllogism ($A \vee B, \neg A \vdash B$) or transitivity ($\Gamma \vdash A; A \vdash B \Rightarrow \Gamma \vdash B$).
With aspect-oriented ontologies this reduction is not necessary, since the mutually contradicting axioms are contextualized. It is only necessary to keep the possible worlds representing the contexts isolated from each other by forbidding accessibility relations between them.

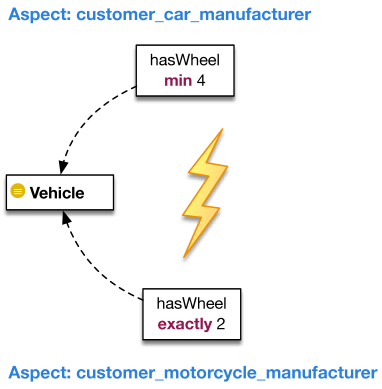

Figure 3: Different axiomatizations of the same concept, according to two different stakeholders.

- Multi-agent belief using doxastic logic
  A further conceivable scenario involving the deployment of semantics similar to that of logic K is the interpretation of contexts as agents' beliefs about axioms and facts in a knowledge base. Doxastic Logic uses a modal operator $\mathcal{B}$, which expresses the belief of an agent in axioms and facts. By indexing the belief operators, we obtain a multi-modal logic with multiple belief operators, each representing the belief of one particular agent. A typical scenario in ontologies is to contextualize factual knowledge with an agent's belief, as described by the *Context Slices* ontology design pattern[4] [Wel10]. The pattern is applicable to Abox facts, and, more precisely, to object property assertions only. Typically, the contextualized fact is represented by a contextualized projection of the two individuals involved, where the projections are related to a context individual.
  With aspect-oriented ontologies, this principle can be extended to any kind of axiom, not only Abox facts. A further advantage of using aspects over the context slices

4 http://ontologydesignpatterns.org/wiki/Submissions:Context_Slices

pattern is that the contextual knowledge (the statements on the belief) are more clearly separated from the actual knowledge. Using the context slices pattern, the original object property assertion does not remain intact. There is only a relation between the projected individuals. Taking the knowledge out of its context and reusing it in another ontology in a different context or not contextualized at all, is difficult. Figure 4 shows how the fact that an agent believes in an object property assertion may be modeled using ontology aspects and the context slices pattern.

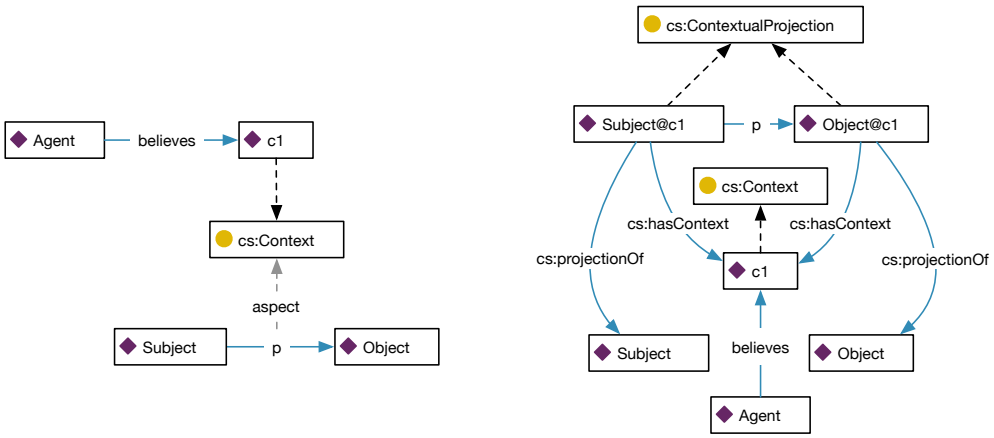

Figure 4: An agent's belief of an Abox axiom represented as an ontology aspect (left) and using the conctext slices ontology design pattern (right). With aspects, the contextualization with agents' beliefs can be applied to any kind of axiom, not only Abox facts.

- Temporal aspects using temporal logic
  While ontologies are supposed to capture universally valid knowledge, there are situations in which an application might make use of knowledge that is contextualized with time. A problem that arises in making universal statements in the form of *A is valid during time interval T* is that actual domain knowledge and contextual (time) knowledge become tangled.
  Figure 5 depicts such a scenario using a temporal aspect on an Abox fact capitalOf(Bonn, Germany), which was true between 1949 and 1990.
  Using temporal aspects, each time instance is interpreted as a possible world, and *after* (and *before*) are accessibility relations, which are reflexive and transitive. The corresponding conditions on the Kripke frames for a temporal logic are:
  - $(M) : \Box A \rightarrow A$
  - $(4) : \Box A \rightarrow \Box\Box A$

  The temporal aspect is then the class expression
  after value 1949 and before value 1990,
  which includes the values 1949 and 1990 due to the reflexivity of the before and after relations.
- Access Rights using Deontic Logic
  Controlling access to sensitive knowledge contained in knowledge bases is important especially in corporate contexts. Currently, access restrictions are part of the technological infrastructure, for example, SPARQL endpoints may be accessible only after authentication.
  With aspects it is possible to model access restrictions in a more fine-grained manner, by providing arbitrarily many access permissions for each axiom or fact in the knowledge base.
  The problem with restricting access to parts of knowledge in a globel model is that agent who are forbitten to access certains parts of the model have to deal with an incomplete model of the world. This may (not necessarily but possibly) lead to inconsistent states in the local world view of such an agent.

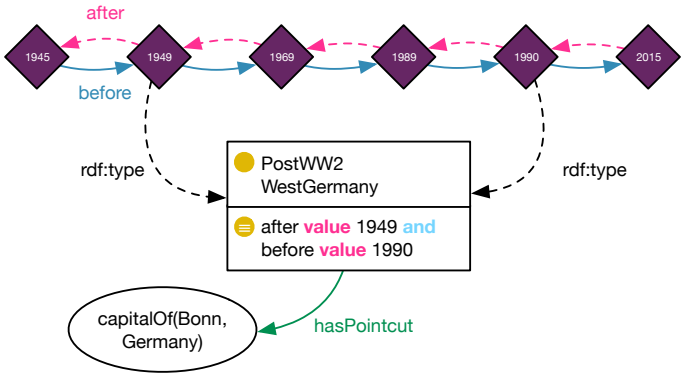

Figure 5: A temporal aspect using temporal logic with a transitive accesibility relation representing the temporal *before* relation (and its inverse *after*).

By modeling access restrictions as aspects using deontic logic, it is possible to model an alternative sub model for an access restricted part of the global model. It is possible to model the "forbidden" state not as failure but rather provide an alternative, simplified truth. This may be done by providing a set of "repair axioms" that complete the world view of the agent in such a way that it becomes consistent again.
Multi standard deontic logic for modelling access permissions over axioms for different agents by having a serial accessibility relation a_i for each agent i in order to obtain the axiom

- $(D) : \Box_i A \rightarrow \Diamond_i A$

2.1.4 Industrial Scenarios

- Development of Multilingual Ontologies
  A recurring challenge in cross-cultural contexts is multilingualism. Examples of situations requiring multilingual assessment of knowledge include global corporations with subsidiaries in different countries that desire to build a company-wide body of knowledge, or business intelligence scenarios addressing international markets. Another example are multilingual web applications with databases of goods that, depending on the counry have different names but also different components or ingredients, like for example a drug database with the same medication having different ingredients due to different law or patent situations in the different countries.
  The challenge in this scenario goes beyond the need to attach multilingual labels to concepts and objects. Instead, the intensional description of entire concepts or objects may differ completely.
  This scenario requires a contextualized conceptualization of the domain of interest, where language or cultural background constitutes the context. It possibly needs to deal with global inconsistencies due to locally consistent but globally contradicting world views.
- Role-Specific Views on Business Processes
  One of the goals of Corporate Smart Content research is the context-sensitive provisioning of information for knowledge-workers. Business process models are considered as one possible source of context-information.
  This scenario is based on the conceptualization of entities relevant in a cross-department business process model in a production environment. The process model vocabulary includes
  - 195 concepts and
  - 52 properties

  These may or may not be relevant in 12 contexts which represent the view of different stakeholders as well as 3 meta contexts representing technical aspects of the process

context.
The contexts are
- Constructor
- Service
- Controller
- Procedure
- Transport
- Component
- Procedure Map
- Request checklist
- Facility
- Offer
- Client
- Documentation
- Inquiry
- Problem
- Glossary
- Context

These context can be modeled independently of the domain concepts and properties and then later combined.

- Street names in location-based search
  A concrete application scenario from the CSC domain is a web application providing locality-based search for services and sightseeing spots for tourists in the city of Berlin [PST+14]. It also involves a search facility for historical information about locations in Berlin including semantic descriptions of the localities, which are, among others, provided by modeling streets and their names.
  One problem that needs to be adressed is that street names in Berlin have undergone a significant amount of change during the last century due to natural growth of the city, but also due to its dynamic history.
  Since the goal of this scenario is change of concepts over time, it is a candidate for the application of temporal aspects. As in the general scenario description for temporal aspects, this scenario does not only include the changing of labels but also attributes of entities (many streets kept their name after the wall was built, but were cut in half by the wall, so that only a part of the original street kept the original name, while the other half became a new street with a new name). Also the conceptualization changes slightly since categories representing historical periods (e.g. "Street in East Berlin") which in themselves only exist during certain periods of time are deployed.
- Access Rights in Corporate Wiki using Deontic Logic
  Corporate Smart Content addresses both content that is considered external and internal for enterprises. Internal content may be contextualized by access restrictions to certain groups of coworkers.
  This scenario covers the above problem in the context of internal corporate wiki systems combined with a knowledge extraction task [PST+14]. It integrates into the authentication and access-rights management system of the wiki.
  Users create content in the wiki. While doing so, they fulfill one or more roles in the enterprise. The content they provide is access-restricted based on the enterprise specific role model.
  A knowledge extraction process creates axioms and facts representing the knowledge stored in the wiki.
  This scenario requires a model which aids in applying the same access restrictions that apply to the textual content to the factual knowledge gained by the extraction process. It is a candidate for the application of the general access rights research scenario.
- Temporal Attribution of Facts in a Corporate Wiki
  This scenario addresses a different problem in the same wiki application described above. Corporate knowledge is dynamic. As users add content to the wiki, knowledge my become stale (i.e., still valid but not useful anymore) or invalid and replaced by new, possibly contradicting knowledge.

This is a use case for using temporal aspects as described in the research scenarios.

2.2 Complex Entity Recognition

Under the term Complex Entity Recognition (CER) we combine a series of technologies from the Natural Language Processing field in order to go a step beyond recognizing simple named entities and extract clusters of related information from text.

2.2.1 Overview

As many other Information Extraction (IE) approaches, CER deals with large amounts of text documents. The CER process however, is divided into 3 phases.

The first phase employs traditional IE steps. These documents are processed in an NLP pipeline where the at first we perform traditional named entity recognition. The second step in our approach is relation extraction, the explicit relationships between entities in the text documents are recognized, extracted and matched to possible candidates in our knowledge base (KB). The named entity recognition and relation extraction steps are part of the reduction phase, at the end of this phase the text documents are reduced to a set of entities and relationships.

The second phase is an enrichment phase, where we add additional information from the KB to the extracted entity/relationship sets. This added background information can be the types of recognized entities (e.g politician, musician, financial_institution, capital_city, etc.), super-types (person, location) or related entities (adding the country for a city, the band name for a musician). The enrichment phase is done with a series of inference rules since adding irrelevant information can skew the usefulness of the mined complex entities.

In the third phase we mine the actual complex entities, this can be achieved trough multiple approaches based on the type of entity we wish to mine. Topic entities can be mined trough an adapted LDA approach, instead of counting the frequencies of words in documents, we count the frequency of entities in the enriched entity sets. In our algorithm we specify the number of complex entities we wish to mine and the output is a cluster of entities that have a high probability of appearing together in the document corpora. These complex entities each describe a topic such as politics, food, chemistry, music, movies, etc. When taking in the time aspect and the evolution over time in the usage of the entities that form a topic-entity, a topic-entity can represent a trend. When looking at the co-occurrences in the enriched entity set, we can mine various implicit relationships. For example a politician might often be mentioned together with people from the art domain, or a fashion designer might often be mentioned together with various musicians.

In the following subsection we will present a couple of industrial and research scenarios where CER has been or can be employed.

2.2.2 Industrial Scenarios

- Trend Monitoring in the Business Domain
  One of the applications that can have a high impact on the corporate bottom line is the monitoring of business and consumer trends. Many companies already use social media monitoring in order to better understand what their customers think of specific brands and products. Understanding the trends that develop in the market can be just as important for companies. In the last years we have observed some major shifts in the financial as well as technological markets. Not only the financial crisis was a trend that can be predicted by trend analysis but also major technological shifts such

as cloud computing, big data or cognitive intelligence. The technologies employed for complex entity recognition, allow companies to automatically gather and analyze very large corpora of relevant information. By analyzing this information with our enriched topic modeling approach, we enable users with the relevant insight into the data to detect such technological and financial shifts in time to make more informed business decisions.

- Detecting Experts in Specific Fields
  Understanding the expertise of employees in a specific field can prove a crucial insight for the company strategy. However, such a task is hard to perform since there is no way to gain this insight without a significant overhead. By analyzing the documents written by an employee we can statistically determine the areas of expertise of an employee. Although not a perfect approach, this can provide managers with a fast insight over the areas of expertise of all employees in a company, and thereby improve decisions regarding assignment of specific tasks.
- Automatic Document Classification
  Companies have always had a problem with finding relevant data, and with the big data age and the growing accumulation of data, these problems have increased exponentially. In many cases companies don't really perform any complex analysis on the data, but they need to search it and classify it in a fast and intuitive manner. However, this is not possible since all this data is thrown into one bucket and no one knows what's really in there. Document classification schemes such as libraries use are impractical and not realistic in a corporate scenario since they would require massive investments of employee time. The only solution is to use machine learning to automatically classify this information. This is one of the application areas of CER, where we use enriched topic modeling in order to automatically derive a structure from documents and classify this information based on that structure. This allows companies to quickly find the information they use based on a specific topic such as: financial data, annual reports, CEO statements, project milestones, press releases etc.
- Trend-based exploration
  In some cases understanding large document collections can be more difficult than simply exploring the collection based on automatically extracted topics. One of the reasons is that topics evolve over time. Looking over a wide enough timespan we can reduce the number of topics to some general topics such as politics, taxes, earnings, events etc. However, if we look in a more recent time span these topics change quite a lot. Things like annual vacation, earning report 2015 etc. are much more important. Being able to explore documents based on recent topics and recent growth in relevance can become an important tool in order to retrieve, browse and understand large document collections.
- CE-based visual exploration
  In many cases understanding the information from the corporate data trove is crucial. In these situations traditional approaches reach their limit, and a more intuitive visual exploration is needed. In our CER approach we reduce the unstructured data companies posses to a knowledge layer, where the data is represented by semantic entities and relationships between these entities. By applying visual graph exploration, new insights can be gained and clusters of information become visible by applying the right algorithms.

2.2.3 Research Scenarios

- CE-based Entity Resolution
  The main way entity resolution is implemented in most systems is based on contextual clues. The context around an entity is considered to be the main source of features in oder to disambiguate the entity, and link it to the correct entry in a database. Another approach deals with the overall topic of the document where an entity is found. So the topic of the document is considered to be one of the major deciding features. The approach we propose combines topic mining with contextual entity linking and the correlation of these two approaches in order to improve the precision of the entity resolution step.

- CE-based Association Extraction
  On of the inherent assumptions is that if multiple entities are present in a mined complex entity, they are strongly interrelated and all of them belong to the same topic. However if a document belongs two multiple topics, then there must also be a relationship between the entities that form those topics. With this approach we can mine associated entities, not by cooccurrence rules but by the adherence to a generative model, and the coocurrence between topic models.
- CE-based Event Detection and Tracking
  LDA has been employed for many EDT approaches. By reducing the input data to a set of knowledge, that is entities and relationships, and then mining complex entities, we can use simple inference rules or pattern mining techniques on the complex entities in order to determine if they are an event or not. By employing graph similarity measures on new mined entities, we can determine if they represent one of the previous mined complex entities, and therefore track the event.

2.3 Semantic Pattern Mining in Event Streams

Extensive and growing data sources like event streams make a manual observation of information flows by humans impossible. In Complex Event Processing (CEP), streams are analyzed and monitored to detect sophisticated situations and phenomena in real-time. Event Processing Languages (EPL) provide real-time stream query capabilities to aggregate and composite new complex events. The to-be-recognized patterns need to be known at first when defining an EPL query but streams supply continuous, varying data with changing patterns over time. EPL queries are not capable of capturing unforeseen patterns which emerge over a long time period.

Pattern Mining solves this arising problem. It aims to discover and learn considerable frequent patterns through interconnections and correlations in databases. Event instance sequences (EIS) are monitored and processed to get frequent behavioural patterns.

Semantic Pattern Mining combines the traditional Pattern Mining with Semantic Web technologies by using structured data stored in ontologies.

In the following some examples how Semantic Pattern Mining can be performed:

1 The semantic enrichment of event streams with background knowledge extends the information content of the primitive events and therefore more sophisticated and complex patterns can be learned.
2 The data stream itself can be represented semantically to attach expressive knowledge.
3 The mining algorithms can be semantic, e.g., using semantic similarity measures to evaluate the distance and semantic relatedness between item nodes in a taxonomy [Res95].

The overall goal is to make sense of enormous amounts of available information and benefit of monitoring and analysis insights.

2.3.1 Overview

Typical sources of events are sensor networks that generate event notifications, e.g., temperature, smoke readings or GPS updates. These can be received periodically, or upon the occurrence of some event, e.g., an incoming order in a restaurant. Sensors monitoring systems may generate primitive events like the status of chemical additive valves, electricity, or the detection of voltage drops from which complex events like system failures or excessive energy consumption can be derived and predicted.

In CEP, complex events are usually expressed in EPL, originally inspired by active databases to process events according to triggers defined as event-condition-action (ECA) [SMMP09]. Based on the approach of detecting state changes automatically and asynchronously, the declarative event specification language SNOOP [CM93] with the event operators for, amongst others, Disjunction, Sequence, and Conjunction for building composite event patterns were introduced.

Based on these operators, EPL executes pattern matching using event operators over incoming event streams to combine multiple primitive events to new, complex ones. Furthermore, a rich set of stream functions can be used to summarize, aggregate and analyze stream data.

Pattern Mining works the other way around. Patterns are unknown at first and learned over time. In traditional databases the detectable patterns stay static and never change. Data streams on the contrary provide new information in real-time and patterns may change frequently. EIS are analyzed according to their occurrence and structure to mine frequent patterns, association rules, time-series, episodes, and many more.

By attaching semantic background knowledge to each event, the information content becomes more exploitable than the raw data. For instance, if a complete ontology graph would be attached the whole data could be accessible and reusable to learn complex patterns.

Traditional Pattern Mining can retrieve the following patterns highlighted in Figure 6. These are split into basic, multidimensional and extended patterns and rules. The subcategories are enlisted and explained in more detail below.

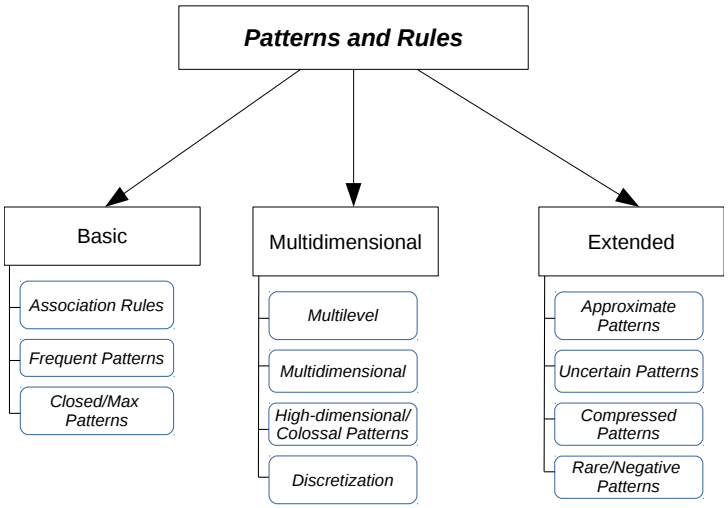

Figure 6: Overview on Patterns and Rules

- Association Rules
  An association rule is an expression of the form $X \rightarrow Y$ with X and Y being sets of items [AIS93][HGN00]. The meaning of such a rule is that whenever a database transaction T contains X then T probably also contains Y. Besides the analysis of market-basket data, association rules are successfully applicable to a wide range of business problems, e.g., decision making, marketing and transaction analysis.
- Frequent Sequential Patterns
  Itemsets, subsequences or substructures in a transaction database are observed with regard to a user-specified minimum support threshold to mine for frequent patterns. Sequential patterns [SL13] represent the correlation between transactions which reflect the most frequent behaviours. Applications are the study of customer behaviour, analysis of mutation patterns in computational biology, or web usage mining. Basic frequent

itemset mining algorithms are Apriori [AS94], FP-growth [HPY00] and Eclat [ZPOL97] [HCXY07].

- Closed/Max Patterns
  In frequent patterns, each sub-pattern is frequent as well. Thus, large patterns contain an exponential number of smaller frequent sub-patterns [HCXY07]. To overcome the possibly huge number of patterns satisfying the minimum support threshold in frequent pattern mining, closed frequent [PBTL99] and maximal frequent [Bay98] pattern mining methods were proposed for the compression of patterns. A closed frequent pattern A is a frequent pattern where there exists no proper super-pattern B that has the same support as A. Closed pattern mining algorithms are for instance A-Close [PBTL99], CLOSET [PHM00] and CHARM [ZH]. A maximal frequent pattern A is a frequent pattern where there exists no frequent super-pattern B such that $A \subset B$. Maximal pattern mining methods are for instance MaxMiner [Bay98] or MAFIA [BCG01].
- Multilevel and Multidimensional Patterns
  Multilevel and multidimensional patterns are mined through multi-level or multiple-dimensional space [CY05]. Multilevel patterns are association rules or sequential patterns according to several levels of abstractions or hierarchies that represent the level of granularity for items. From multidimensional information, e.g., customer purchase sequences associated with region, time and customer group, complex frequent sequential patterns can be mined. To find sequential patterns from *d*-dimensional sequence data algorithms like AprioriMD and PrefixMDSpan [YC05] were developed. Possible usage is the mining of patterns from temporal-spatial data.
- High-dimensional/Colossal Patterns
  High-dimensional datasets are characterized by small number of transactions and large number of items in each transaction which results in increased running time and large result sets with only small and mid-size frequent patterns. Sohrabi and Barforoush [SB12] provide efficient algorithms and pruning rules for colossal pattern mining based on vertical format. Pan et al. proposed CARPENTER [PCT+03] and COBBLER [PTCX04], two methods frequent closed itemsets by integrating vertical format and row and column enumeration.
- Discretization
  Applying pattern mining techniques to real-time streaming data consisting of discrete and continuous data instead of static data, problems may arise [CY05]. Discretization is the process of transforming a discrete or continuous variable, such that it takes fewer number of values. For instance, clustering or time window approaches may transform continuous data sequences into discretized, abstracted representations. The resolution of discretization refers to the extent of intervals to not miss potential discoveries for meaningful patterns [Bay00]. Statistical approaches include univariate and multivariate discretization.
- Approximate and Uncertain Patterns
  Approximate frequent itemsets (AFI) [LPW+05] aim to limit the number of random noise and errors in frequent itemsets due to their strict definition of support. Liu et al. developed an algorithm to mine AFIs. The definition of support for these kind of patterns is more relaxed and allows some degree of error.
  Mining uncertain data for frequent pattern mining is a known problem. To handle this uncertainty, items in a transaction are assigned a non-zero probability. For this assumption the deterministic procedure to determine frequent itemsets is impossible, and needs to be adapted to probabilities. Aggrawal et al. [ALWW09] study this problem and extend broad classes of algorithms to the uncertain data setting.
- Compressed Patterns
  As mentioned earlier, the number of discovered frequent patterns can explode with a low minimum support value. One approach to overcome this issue is the compressing of frequent pattern sets, e.g., by finding a concise representation describing the whole collection of discovered patterns. Major approaches are the *closed frequent itemsets* and *maximal frequent itemset*. Xin et al. [XHYC05] study the problem of similarity between patterns and how to define and discover qualitative clusters for compression techniques. Lam et al. [LMFC14] propose two algorithms, SeqKrimp and GoKrimp, for

mining compressing patterns.

- Rare/negative Patterns
  As opposed to frequent itemsets which cover frequent behaviours and trends, rare and contrasting itemsets can also be of high interest, e.g., to discover rare combinations of symptoms in medical databases [SNV07]. Szathmary et al. propose the algorithms Apriori-Rare, MRG-Exp and Arima.

2.3.2 Industrial Scenarios

According to Luckham [Luc12] and Agrawal et al. [ADGI08] the following typical industrial scenarios for Pattern Mining and CEP evolved.

- Financial Systems and Services
  High-performance event processing in stock and financial front-office trading was one of the earliest areas in which CEP platforms were used for real-time processing. Tens of thousands of events per second are processed in real-time to automatically make buy or sell decisions according to live updates in stocks. Complex and proprietary algorithms are used to detect favourable trading situations between multiple market data feeds.
  The monitoring of events is the basic application, but still very helpful. The pattern-based market analysis of stock trading events can detect new trends, customer behaviour is tracked to compute credit ratings in real-time, violations of company policies and frauds can be detected automatically.
- Fraud Detection and Security
  Fraud-detection systems make use of rules triggered by specific event patterns of fraudulent behaviours, e.g., to track valid paths of shipments and detect anomalies such as food contamination in supply chains. Automated event processing is most effective for the detection of misuse of stolen credit cards. The area of fraud-detecting is expanding and being improved frequently and constitutes a positive trend toward building standard sets of CEP rules. In contrast to fraud-detection, security is about preventing problems from happening instead of their detection. CEP can help detecting event patterns of illegitimate traffic by continuously monitoring the event traffic.
- Transportation
  Airline, trucking, railway or shipping transportation use CEP techniques for monitoring of maintenance schedules and inventories. Furthermore, transportation delivery tracking and coordination of for instance truck fleets can be improved.
- Business Process Management (BPM) and Business Process Mining
  The actual execution of activities in an enterprise is reported in a form of (complex) events by running processes for analysis of process behaviour in real-time. The integration of CEP and BPM provides knowledge derivation during execution time, favourable for control over pure system interactions, unwished situations or observation of current trends.
- Energy
  Electricity grid control, or smart grids, consist of a power transmission grid coupled with a two-way event driven monitoring and control grid. Up-to-date examples are consumers scheduling power consumption of their smart home, or arbitrary process control in factories.
- Health Care
  Even though lots of applications for health care CEP have been developed in the past, this is a field of ongoing innovation. Examples are event-driven systems for monitoring patient treatments in hospitals, hygiene compliance reminder for healthcare workers and reporting and processing the patient status events to facilitate the communication of health care facilities.

2.3.3 Research Scenarios

Figure 7 provides an overview of current research scenarios for Semantic Pattern Mining. We differentiate general mining methods from applications and extensions. The Mining methods are split up into the basic methods, interesting patterns and distributed, parallel and incremental patterns. Furthermore, we highlight extended patterns and pattern-based applications. The subcategories are enlisted and explained in more detail below.

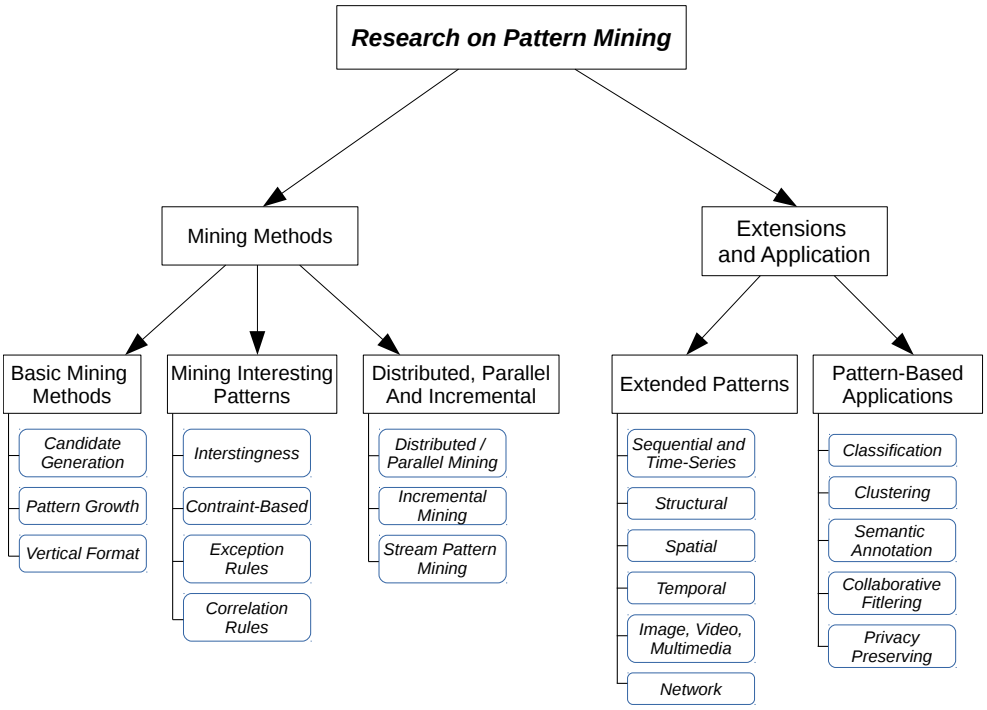

Figure 7: Research Overview of Pattern Mining

- Basic Pattern Mining
  - Basic Pattern Mining Methods [AH14] include candidate generation (e.g., Apriori algorithm, portioning or sampling), Pattern Growth (e.g., FP-growth, HMine, FPMax, Closet+) and Vertical Format (e.g., EClat, CHARM). Candidate generation iteratively creates all frequent patterns, called *candidate patterns*, of length k based on a minimum support threshold in each scan of the database. To overcome the large number of candidates, especially for long patterns, and multiple scans of the database the approach of Pattern Growth was introduced. It mines frequent patterns without candidate generation, but with divide-and-conquer methodology to reduce the search space and compute with high performance. The afore-mentioned methods mine frequent itemsets in horizontal format which means that the database is scanned multiple times. Vertical format [mGjW10] on the contrary scans the database only once to get the itemsets which are then conjugated iteratively to obtain the candidate sets.
- Mining Interesting Patterns
  - Interestingness

    To assess object relationships within data patterns, interestingness measures have been proposed in statistics and data mining. According to Geng and Hamilton [GH06] the interestingness of a pattern is determined by means of the following criteria: *conciseness, coverage, reliability, peculiarity, diversity, novelty, surprisingness, utility* and *actionability*. These nine criteria can be further classified into *objective* and *subjective*. Objective measures are based only on the raw data without any knowledge about the user or application. A subjective measure on the other hand takes into

account both the data and the further required knowledge about the user and the domain. Semantic-based measures are considered a special type of subjective measures and take into account the semantics and explanations of patterns and involve domain knowledge from the user.

- Constraint-Based Pattern Mining
  The aim of Constraint-Based Pattern Mining [NZ14] is to find different types of patterns satisfying specified constraints, e.g., in frequent itemset mining minimum support constraints, or in association rule mining minimum confidence constraints are used.
- Exception and Correlation Rules
  According to Hussain et al. [HLSL00] mining interesting rules is one of the important data mining tasks since they bring novel knowledge for advantageous actions. They defined exceptions as rules that contradict the common belief of a user. Since exceptions are usually a minority, e.g., not yet known or omitted, they form interesting rules to be estimated from the mined rules.
  Association or correlation rules and patterns recognize interesting event relationships with strong correlation.

- Distributed, Parallel and Incremental Mining
  - Distributed and Parallel
    The rapidity of pattern discovery in increasing data collections will always be essential. In huge datasets with sequences and corresponding subsequences, sequential pattern mining will generate an explosive number of frequent subsequences for long patterns. El-Hajj and Zaiane [EHZ06] develop a parallel frequent mining algorithms that generates the set of maximal patterns and fits well a distributed or cluster environment to obtain scalability in pattern discovery. Another approach is Par-CSP [CHP05] which is a Parallel Closed Sequential Pattern Mining algorithm for distributed memory systems.
  - Incremental Pattern Mining
    Incremental databases comprise frequent operations like additions, deletions and modifications. In actual research, incremental updating tree data structures are developed to represent the database transactions and updates efficiently using currently available memory, e.g., incremental prefix-tree structures such as CanTree [LKLH06], CP-tree [TAJL08], etc.
  - Stream Patterns
    While traditional databases store static data and patterns of interest do not change significantly over time, streams with their high volume and velocity require real-time or near real-time processing due to the volatility of the incoming observations, which can be stored for a limited time only. Mining real-world time series and streaming data creates a need for new technologies and algorithms, which are still being developed and tested by data scientists worldwide. Finding the most efficient representation of streaming data, developing privacy-preserving methods for data stream mining, is an actual challenge. Proposed algorithms to mine pattern in data streams are, amongst others, Lossy Counting [MM02] and DFPM [YCL+06] [JA07].
- Extended Patterns
  - Sequential and Time-Series
    As denoted earlier, sequential pattern mining is the discovery of frequently occurring ordered events or subsequences as patterns. Han [Han05] outlines a wide overview on time-series and sequence data. A time-series database consists of sequences of values or events obtained over repeated measurements of time. Applications are, e.g., stock market analysis, economic sales forecasting, inventory studies, and many more. Nowadays, the amount of time-series data is increasing rapidly, thus the arising problem of how to find correlation relationships, regular patterns and trends in time-series data arises. Current challenges include mining entity-related time series and incremental pre-processing of continuous time series and data streams in parallel to the data mining process.
  - Structural, Spatial, Temporal and Multimedia Patterns
    Some further extended data type patterns are structural patterns which rely on graph-, tree- or lattice- like data structures. GraphMiner [WWZ+05] is a research

prototype system for mining frequent patterns from large, disk-based graph databases. Spatial data may be mined for co-location and temporal data my be mined to detect evolutionary and periodic patterns. Giannotti et al. [GNPP07] introduce trajectory patterns to mine spatial regions frequently visited using large spatio-temporal datasets. The discovery of useful knowledge for decision making from unstructured multimedia databases, e.g., images or videos, is also a challenging problem.

- Network
  Mining network patterns, like roles in networks [RA15], can detect node-level connectivity patterns. Roles can be mined by analyzing the graph-like structure of events built upon their relations and correlations. Two nodes belong to the same role if they are similarly structured. For instance, SpaceROAM [HTP15] was designed to discover unknown role patterns in collaborative systems based on domain-dependant features.

- Applications
  - Pattern-based Classification
    Pattern-based classification can be used to extract more accurate and interpretable models from data describing important data classes or to predict categorical labels. An overview over this topic is provided in [BNZ11] covering topics like class-sensitive patterns, model-(in)-dependent pattern selection and post-processing.
  - Pattern-based Clustering
    In pattern-based clustering a set of data objects form a pattern-based cluster if these objects follow a similar pattern in a subset of dimension [PZC+03]. This approach is flexible in providing interesting and important insights in applications where conventional clustering methods may meet difficulties.
  - Semantic Annotation Semantic annotation of patterns includes semantic background knowledge to mine for interesting patterns. Semantic Complex Event Processing (SCEP) uses semantic background knowledge to map generalized CEP patterns to heterogeneous event streams for processing. Event Pattern Mining enriches (or annotates) event streams or logs with semantic background knowledge to learn CEP patterns and generalize them inductively. Multi-dimensional complex event patterns and their generalization can be learned in Semantic Pattern Mining to detect meaningful events from these in SCEP. For instance, in [LFTP15] real-time news streams are enriched with background knowledge and mapped to semantically defined CEP patterns. Semantic Pattern Mining can improve this approach by automatically mining frequent news patterns and adapting these over time.
  - Collaborative Filtering
    Pattern-based collaborative filtering can be used to predict the customers' behaviour by clustering customers and deriving sequential patterns among items for each cluster in each time period [CYKS12].
  - Privacy Preserving
    Privacy Preserving is an important concept nowadays and aims to hold the balance between hiding restrictive patterns and disclosing non-restrictive ones [OZ02]. The main procedure, also called sanitization process, is to hide a group of frequent patterns which contain highly sensitive knowledge.

3
Test Datasets

This section describes extracted or generated datasets suitable for evaluating custom approaches in Aspect-oriented Knowledge Engineering, Complex Entity Recognition and Semantic Pattern Mining.

3.1
Aspect-oriented Knowledge Engineering

Aspect-oriented ontology development tackles the problem of modeling contextualized knowledge. One of the main goals of aspect-oriented ontology developmentis to provide methdodological support for modeling the actual domain knowledge and the context information independently of each other.

The hypothesis behind this is the assumption that this separation of concerns leads to a better modularization of the developed ontologies. As pointed out in [PS09], better modularization is helpful in

- improvement of reasoning and query result retrieval performance
- scalability for ontology evolution
- maintenance
- complexity management
- amelioration of understandability
- reuse
- context-awareness
- personalization

Datasets for evaluating aspect-oriented ontology engineering approaches are therefore ontologies with a measurably good modularization.

In order to define what a "good modularization" is, we resort to existing quality metrics that have been established in the field of ontology engineering and, in particular, in ontology modularization. These metrics have been inspired by modularization metrics from software engineering. A seminal work in evaluating the quality of ontology modularizations is [ED13]. The metrics they propose are *cohesion* and *coupling*.

Cohesion quantifies the degree in which an ontology module is internally connected. The more concepts in an ontology module reference other concepts in the same module, the higher the cohesion value.

Coupling quantifies how much different ontology modules are connected to each other. The more concepts in an ontology module reference other concepts in another module, the higher the coupling value.

Cohesion and coupling metrics are normalized to yield values between and including 0 and 1.

A "good" ontology module is considered as an ontology module that has a high cohesion and a low coupling value.

Table 1 provides the cohesion and coupling values for the LUBM ontology[5], the VICODI ontology[6] and the DBpedia ontology[7].

At the time of writing this study, ready-to-use implementations of cohesion and coupling metrics are not available. As a first step it is necessary to implement these metrics and

5 http://swat.cse.lehigh.edu/projects/lubm
6 http://cordis.europa.eu/result/rcn/34582_en.html
7 http://wiki.dbpedia.org/services-resources/ontology

apply them to all candidate gold-standard and to all evaluation candiate (aspect-oriented) ontologies.

The selection of candidates as gold-standards therefore needs to be done based on different metrics, such as ontology size, number of modules, if an ontology exists in modularized form (distributed accross different files and reused by using owl:import directives) as well as the detection of candidates for contextualized knowledge as described in the application scenarios.

Appendix 6 provides an exhaustive list of gold-standard candidate ontologies.

Dataset	Ontology design	Module	COH	COP	|Loc|	COH_{W-Avg}	COP_Des
LUBM	Monolithic design	Univ-Bench	0.0566	0	68	–	0
	Modular design 1	OM-Publication	0.0984	0.0666	15	–	0.0245
		OM-Person-Organization	0.0851	0.0125	53	–	
	Modular design 2	OM-Publication	0.0984	0.0666	15	–	
		OM-Person	0.1102	0.2051	39	–	0.16596
		OM-Organization	0.0879	0.1632	14	–	
		OM-Organization; OM-Person	–	–	14; 39	0.1043	
VICODI	Monolithic design	VICODI	0.03044	0	194	–	0
	Modular design 1	Module 0	0.1003	0.3333	30	–	
		Module 1	0.0606	0.6666	40	–	
		Module 0; Module 7	0.1003; 0.0634	–	30; 22	0.08477	
		Module 2; Module 3; Module 8	0.202; 0.2666; 0.01111	–	12; 5; 25	0.0960	0.5257
		Module 0; Module 1	0.202; 0.266; 0.0111	–	30; 40	0.08053	
	Modular design 2	Module 0; Module 7; Module 9	0.1341; 0.4444; 0.333	–	22; 3; 4	0.1937	
		Module 1	0.0920	0.6666	21	–	
		Module 0; Module 19	0.1341; 0.0634	–	22; 22	0.0988	0.58169
		Module 2; Module 12; Module 18; Module 20	0.1676; 0.266; 0.266; 0.222	–	19; 5; 5; 6	0.2052	
		Module 1; Module 6	0.0920; 0.444	–	21; 3	0.1366	
DBpedia	Monolithic design	DBPedia	0.0025	0	861	–	0
	Modular design 1	Module 0; Module 4; Module 10	0.0025; 0.2222; 0.037	–	674; 3; 28	0.00488	–
		Module 0	0.0025	0	674	–	
		Module 1	0.1538	666	13	–	0.0436
		Module 0; Module 7	0.0025; 0.0888	–	674; 6	0.0033	
	Modular design 2	Person-Module	0.0061	63	204	–	
		Person-Module; Org-Module; Place-Module	0.0061; 0.0089; 0.022	–	204; 192; 49	0.00914	0.04788
		Person-Module; Work-Module	0.0061; 0.0318	–	204; 65	0.01234	
		Org-Module	0.0089	0.0243	192	–	

Table 1: Cohesion and coupling values for LUBM, VICODI and DBpedia from Ensan and Du [ED13].

3.2 Complex Entity Recognition

Table 2 provides information about the datasets suitable for Complex Entity Recognition. We provide fundamental information like the name, domain, format, size, and information about the prominence in the scientific literature by counting the number of mentions in google scholar.

To ensure a common understanding, we explain the semantics of the columns below:

Name	The name of the dataset.
Domain	The domain captured within the dataset.
Format	The format of data, unstructured text, structured or semi-structured.
License	The license of the dataset. Research means that it is only available for research or teaching purposed.
Size	The file size or the number of words
Used for	Areas where the dataset can be employed in. Mostly NLP datasets, some can be used for knowledge base population also
Mentions	Number of hits in google scholar
Language	The language of the dataset

Name	Domain	Format	License	Nr. Words/Size	Used For	Mentions	Language
Reuters-21578 [Lew97]	News	Text	Research	28 MB compressed	NLP	5,340	English
Reuters RCV1[LYRL04]	News	Text	Research	2.5 GB compressed	NLP	158	English
Reuters RCV2 [LYRL04]	News	Text	Research	487K Articles	NLP	5,340	Multilingual
Reuters TRC2	News	Text	Research	2.9 GB compressed	NLP	51	English
20 Newsgroups [Lan95]	Usenet	Text	Research	44 MB compressed	NLP	3,160	English
Common Crawl [Gre11]	Web	Mixed	N/A	151 TB	NLP	367	Multilingual
Pubmed	Biomedical	Text/PDF	N/A	25 Mil Citations	NLP	5,970,000	Multilingual
ClueWeb	Web	Mixed	Research	5TB	NLP	588	English
DeReKo [KBKW10]	Mixed	Text	Research	28 Billion Words	NLP	721	German
DBpedia [ABK+07]	Multi Domain	Structured Data	CC-BY	150 GB compressed	KB/NLP	14,900	Multilingual
Freebase [BEP+08]	Multi Domain	Structured Data	CC-BY	22 GB compressed	KB/NLP	5,530	Multilingual
Wikidata [VK14]	Multi Domain	Structured Data	CC-BY	6.5 GB compressed	KB/NLP	1,580	Multilingual
Framenet [BFL98]	Multi Domain	Structured Data	Research/ Commercial	N/A	KB/NLP	10,700	English
Propbank [KP03]	Multi Domain	Semi Structured Data	CC-BY	N/A	KB/NLP	3,530	English
Nombank [MRM+04]	Multi Domain	Semi Structured Data	N/A	46 MB compressed	KB/NLP	971	English

Table 2: CER Datasets

In addition to the table, we provide a brief description of each dataset to explain its potential.

3.2.1 News Datasets

- The Reuters-21578 is probably the most widely used dataset in NLP research. This document corpus is formed of Reuters newswire articles that appeared in 1987 and have been made available for research purposes by Reuters, free of charge.
- In 2000 Reuters made further datasets available such as RCV1 which is an English collection of news articles from 1996-08-20 to 1997-08-19.
- RCV2 is another dataset released in 2005 by Reuters, but this one is particularly interesting since it is multilingual and covers the same time period as RCV1. It contains over 487,000 news articles in thirteen languages, among which also German is contained. The data can be used to extract and analyze complex entities. Furthermore, inter-language correlations between complex entities can be found and analyzed.
- A newer collection published by Reuters is TRC2, composed of over 1,8 Million news articles, it covers a time period from from 2008-01-01 to 2009-02-28.

The news datasets can be used to extract and analyze complex entities. One interesting aspect is the time-based annotation of these documents. Since they are collected over a time period, we can observe how the complex entities or entity-topics evolve over time and observe which trends manifest in the data. One interesting dataset is RCV2 which contains multilingual articles over the same time period, this dataset could allow us to find interesting correlations between articles in multiple language that contain similar complex entities.

Other news datasets can be generated by crawling various news sites. For example we could crawl Google News for all articles containing a specific keyword in a specific time interval. This would allow us to create a news corpus focused on one specific topic, and then to analyze the subtopics resulting by applying the CER approaches.

3.2.2 Crawling Datasets

These datasets are formed by crawling the entire web or a subset of it and collecting a huge amount of text data. These datasets are one of the best examples of Big Data and can be used for testing the scalability of the CER system implementation. Furthermore they can be used to relation extraction and fact extraction, due to the wide domain coverage and abundance of a multitude of sources for the same factual information.

There are 2 widely used Crawl Corpora.

- The ClueWeb dataset is the crawl corpora most widely used for research purposes. It is intensively used by multiple TREC challenges and is subdivided into 2 main releases

  ClueWeb09 which contains over 1 Billion pages in 10 languages. The uncompressed dataset reaches 25 TB in size . This dataset was generated between January and February 2009

  ClueWeb12 is a newer ClueWeb dataset but it's scope is a bit different than that of ClueWeb09. It consists of 733 Million pages, but all of them are in English. The size of this dataset is the same as that of ClueWeb09.
- CommonCrawl is an NonProfit organization that tries to crawl the entire Internet and generates periodic dumps, up to 4 times a year. The size of this dataset is huge, reaching 151 TB uncompressed for the November 2015 dataset.

3.2.3 Enrichment Datasets

These datasets provide us with the background information we require in order to enrich the set of entities extracted from the documents.

- DBpedia
  DBpedia is the most widely cited and most used KB in a research context. It extracts structured data from wikipedia while mapping this data to it's own crowdsourced ontology. This allows us to query the information within Wikipedia wile still using a single unified schema. While being available in all languages present in Wikipedia, the current English release alone contains close to 1 Billion triples.
- Freebase
  Freebase is the KB that used to be made available by Google, but has been shut down in 2015 and no new releases are available. The last dump contained about 2.4 Billion triples, and was therefore a lot smaller than the cumulated size of the entire DBpedia dump. One of the interesting aspects of Freebase is it's high data quality achieved trough intensive manual curation. The quality of the data and the schema allow it to be used for applications that require a higher degree of precision. In the case of complex entity recognition however, this aspect does not have such a high impact.
- Wikidata
  Wikidata is a newer project from the Wikimedia Foundation which aims to be a Wikipedia for structured data. Unlike Wikipedia, the data in Wikidata is entered directly in a structured format and is therefore much more precise and can be queried according to a predefined schema. The current Wikidata size is with it's 4 GB in size quite small. All the data in Wikidata is subsumed by the newer DBpedia dumps and aligned with the DBpedia ontology.
- Framenet
  Framenet is a rather large corpus of sentences annotated according to the theory of frame semantics [Fil82]. A semantic frame is a coherent structure of concepts that describe an event or action and the actors/participants involved in it. By employing automatic semantic role labeling[GJ02] we can identify Framenet frames in our text documents and use them as background knowledge.
- PropBank and Nombank
  PropBank and Nombank are corpora of prepositions annotated with semantic roles, similar to Framenet.

3.3 Semantic Pattern Mining

As explained in Section 2.3, Semantic Pattern Mining aims to detect meaningful patterns by combining Pattern Mining and Semantic Web techniques. The input dataset forms the foundation for developing and evaluating procedures. The necessity for semantically represented data, or at least data that has the potential to be semantically enriched or mapped, is an explicit requirement to perform Semantic Pattern Mining.

This section provides an overview of datasets that are in general suitable for Semantic Pattern Mining. Metadata concerning each dataset is collected and aggregated to provide a wide summary.

3.3.1 Datasets for Semantic Pattern Mining

Table 3 and 4 provides information about the datasets suitable for Semantic Pattern Mining. We provide fundamental information like the name, domain, format, size, but also expressive

information about the content of each dataset like the number of domain concepts, event level or stream distribution[8].

To ensure a common understanding, we explain the semantics of the columns below:

Name	The name of the dataset.
Domain	The domain captured within the dataset.
Format	The file format.
Size	The file size.
Number of Records	The approximate number of records if not directly available.
Dimensions	The number of fields for each record. In CSV or TSV format this substitutes the number of columns, in XML the dimensions are the number of different elements plus number of different attributes, and in TTL it corresponds to the number of RDF attributes.
Numerical (Units)	The units of numerical quantities presented.
Categorical	The nominal values presented. These values are usually textual categories, like codes, groups, names and many more.
Domain Concepts	The number of domain classes and properties. A domain concept is a concept associated to the concrete domain, e.g., treatment procedures in a hospital would be domain concepts, while longitude and latitude values are more of fundamental, geographical concepts. The values values presented here would be needed for a minimal ontology that captures the semantics of the data.
Event level	The expressiveness of the events in the dataset. The event level can substitute to the nominal values of *low*, *medium* and *high*. Low level events are raw sensor update values with no additional information. The raw data is not yet interpreted. Medium level events correspond to atomic domain events. A high level event is a complex event made up of a number of medium-level events.
Distributed Streams	The number of distributed streams for the datasets. This field either displays the number of available separate streams for the dataset, or the possible manual separation into a number of streams.
Annotated	This field holds the information if the dataset is already annotated.

8 The dataset name links to the webpage where this dataset can be downloaded from.

Name	Domain	Format	Size	Number of Records	Dimensions	Numerical	Categorical	Domain Concepts	Event Level	Distributed Streams	Annotated
DEBS 2015	Transportation	CSV	31 GB	1.00E+07	17	dateTime, time, miles, currency, lon, lat	medallion, license, payment type	~10	Medium	-	no
DEBS 2016	Social Network	CSV	150 MB	1,700,000	25	datetime	-	4	-	-	no
BPIC 2011	Health, Hospital	XES	85 MB	150,000	127	age, dateTime, several codes	group, concept:name, diagnosis	~5	Medium	-	no
BPIC 2015	Municipalities	XES	171 MB	263,000	12	dateTime	activity name	-	Medium	5 municipalities	no
City Pulse Traffic Pollution	Traffic Pollution	TTL	35 GB	450,000,000	8	degree (latitude, longitude), dateTime, ppmv	nodeName	~3	Low / Medium	-	yes
GDELT	News, Daily Events	TSV	100 GB	>200,000,000	~60	degree (latitude, longitude), dateTime	country, type, group, ethics, religions and events, news article	~2050	Medium	-	no
Thematic Event Dataset	Thematic Events	TSV	3 MB	15,000	11	-	car brand, country, continent, type, category	many	Low	-	no
Groceries	Grocery Products	CSV	489 kB	9835	-	-	product names	-	Low	-	no

Table 3: Overview for Semantic Pattern Mining

Name	Domain	Format	Size	Number of Records	Dimensions	Numerical	Categorical	Domain Concepts	Event Level	Distributed Streams	Annotated
Movielens 20 M	Movie Ratings	CSV	552 MB	20,000,000	7	rating (1-5), dateTime	movies, genres	27300	Medium		no
Movie Tweetings	Movie Ratings from Twitter tweets	::SV	8.7 MB	493,340	7	rating (1-5), dateTime	movies, genres	24600	Medium		no
Reddit Pizza Requests	Pizza requests on Reddit	JSON	16.4 MB	5,671	33	time in days, timestamp, counts	country, village, request	?	Medium		no
Flickr	Flickr image posts	XML, CSV	20 GB	5,671	105938	dateTime, timestamp, counts	locality, county, region, country, raw tag-name, pool title, label	4	Medium		no
enwiki	Wikipedia revisions	XML	>3 TB	many	14	?	categories, textdata, main	many	Medium		no
Gowalla	Location-based Social Network	TSV	410 MB	6,442,890	6	degree (latitude, longitude), timestamp	-	-	Medium		no
Iris	Iris plants	CSV	10 kB	150	5	cm (petal/sepal width and length)	Iris class	3	Medium	3	no
Plants	Plants growing location	CSV	1.7 MB	34781	potentially many (list of states)	-	plant names, states	22632	Medium	22632	no

Table 4: Overview for Semantic Pattern Mining Cont.

Table 5 provides possible mappings from ontologies to datasets to obtain semantic information. This overview does not intent to provide a complete and static mapping to semantic background knowledge. It should serve as a generic starting point for semantic relations in the datasets. The ontologies acronym with full name, size, format and the mapping to the afore-described datasets can be found in the table.

Background Knowledge	Description	Size	Format	Dataset Name
ONTOSEM	Ontological Semantics	3.5 MB	OWL	BPIC 2011, BPIC 2015
SUBJECT		32.2 kB	OWL	BPIC 2011
QUDT	Quantities, Units, Dimensions and Types Ontology	140 kB	OWL, CSV, RDF/XML	BPIC 2011
UMLS	Unified Medical Language System	88.2 kB	-	BPIC 2011
MeSH	Medical Subject Headings	658 MB	RDF/TTL, CSV	BPIC 2011
DBpedia	DBpedia Ontology	2.4 MB	OWL, NT	City Pulse, GDELT, Groceries, Movielens 20 M, Reddit Pizza Requests, Flickr, enwiki, Gowalla, MovieTweetings
GeoNames	GeoNames Ontology	533.8 kB	RDF	City Pulse
NCBI	National Center for Biotechnology Information (NCBI) Organismal Classification	478 MB	RDF/TTL, CSV	Iris, Plants
MO	Movie Ontology	113 kB	OWL	MovieLens 20 M, MovieTweetings

Table 5: Ontology Dataset Mapping

3.3.2 Details on the Datasets

Some of the afore-mentioned datasets have been released as the basis for conference challenges that usually take place each year.

For instance, the DEBS (Distributed Event-Based Systems) Grand Challenges frequently provide datasets consisting of data from sensors as events. With the help of Semantic Pattern Mining, complex information about frequent items and paths, time-aware situations, location data, strategy and procedural information can be discovered.

BPIC (Business Process Intelligence Challenge) provides participants with real-live process event logs to obtain insights into the captured processes. Underlying models and execution paths can be mined, and complex and semantically enriched roles be discovered. Usually such logs are used for general Process Mining where cases and activities are accumulated and observed to gather a better overview. Additional textual information in element descriptions can be enriched with semantics to understand the procedures or activities that take place. The timely observation of process logs can bring rich insight into process flows.

The CityPulse Dataset on traffic pollution is the only dataset that is already annotated with semantic information. This dataset could be used for Semantic Pattern Mining on semantically enriched event streams. CityPulse hosts several datasets that are build upon a

number of ontologies including SAO ontology, MUO ontology, UCUM vocabulary, Timeline ontology, SSN ontology and City Traffic ontology [TBA+14]. All datasets are represented in Turtle (TTL) format.

The Global Data on Events, Language, and Tone (GDELT) database monitors the world's broadcast, print and web news from all over the world. It contains over a quarter-billion records that are organized into tab-delimited files by date. Many of the entries are mapped to codes from the CAMEO event taxonomy[9].

The Thematic Event Dataset is built upon thematic tagging of events and subscriptions instead of semantic models such as ontologies. Hasan and Currey [HC15] define *thingsonomies*[10] which are thematic tags that describe the themes of types, attributes, and values and clarify their meanings. Their hypothesis is that extra tags for subsriptions and events improve effectiveness and time efficiency in heterogeneous environments and domain-specific knowledge exchange.

The Groceries dataset contains a collection of receipts with each line representing one receipt and the items purchased in a grocery store. Usually applied for Association Rules, this retail based application can be utilized for Semantic Pattern Mining to detect semantically equal frequent items or complex correlations based on background knowledge.

With the help of item ratings in the MovieLens 20 M dataset, unstructured review text, movie titles and genres can be semantically analyzed and enriched to understand interests and sentiments, make complex recommendations, mine composite item relations or perform temporal data mining. The MovieTweetings dataset consists of ratings on movies obtained from public and well-structured Twitter tweets [DDPM13]. This dataset is modeled after the MovieLens dataset to make them as interchangeable as possible.

The Reddit Pizza Requests dataset contains a collection of textual pizza requests from the Reddit community "Random Acts of Pizza". Altoff et al. [ADJ14] provided the dataset and studied the role of social and linguistic factors when asking for a favor.

The Flickr dataset [ML12] usually used for image classification and labeling consists of manually annotated image collections from Flickr. The large amount of available metadata, e.g., the location the photo was taken, the associated tag names or the pool title the image was shared in, may be used for semantic enrichment.

The complete Wikipedia edit history (up to January 2008) is captured in the enwiki dataset. It contains processed metadata for all revisions of all articles extracted from the full Wikipedia XML dump as of 2008-01-03. This dataset decompresses to several Terabytes of text containing information about the revision like the list of categories, images, cross-references and external links. It therefore holds high volume data and many starting points for semantic enrichment to perform Semantic Pattern Mining.

Gowalla [CML11], a location-based social networking website, shares users locations by checking-in. Consisting of check-in time, location and a friendship network, complex location-based semantic patterns can be mined.

The two machine learning datasets Iris and Plants from the domain of biology hold information about several plants with some characteristics and the growing location area. Complex patterns concerning the growing, whether conditions, humidity etc. can be learned by incorporating of semantics.

In contrast to the described datasets, several synthetic event data benchmarks are considered relevant. In the literature, lots of researches draw on synthetic datasets based on several data generators developed because of their dynamic and adaptable nature. Synthetic benchmarks and datasets worth mentioning are the Big Data Benchmarks and Transaction Processing Performance Council (TPC) Benchmarks. The IBM synthetic dataset generator [AS94] is widely used in pattern mining research.

9 The GDELT Project, last visited April 2016
10 Thingsonomy = thing and taxonomy

4
Summary and Outlook

Once an innovative idea is developed, the underlying algorithm, framework or technology needs to be evaluated to prove its novelty, problem-solving capability, usability, performance and other measures.

The basis for evaluation is a suitable real-life or synthetically generated dataset with the capability to mine for complex semantic patterns. In traditional Pattern Mining all kinds of datasets, e.g., from the machine learning area, can be used. Extensive data points containing statistical information and measurements may suit perfectly for regression or correlation tasks, but unfortnuately most of them do not contain (m)any domain concepts which have the potential to be semantically enriched. Thus, the number of datasets applicable for *Semantic* Pattern Mining is rather small. The datasets in Section 3.3.1 were selected and listed due to their possibility to 1) be enriched with semantic background knowledge and 2) to mine for complex semantic patterns.

In the semantic web, data is represented in a structured way to ensure interoperability and exchangeability of information. The web of linked data enables people to create data stores like ontologies, build vocabularies and write rules for handling data. Besides the ontologies presented in Figure 5 extensive ontology repositories exist which hold additional metadata:

- BioPortal - Repository for biomedical ontologies
- Linked Open Vocabularies (LOV) - Set of several hundreds of vocabularies for the definition of classes, properties and links in linked data.
- datahub - Powerful data management platform from the Open Knowledge Foundation to search for structured datasets published in Linked Data format, register published datasets, and create and manage groups of datasets. The Linking Open Data Cloud Diagram available at `http://lod-cloud.net/`[11] provides an overview over several datasets published on datahub.

In the traditional pattern mining literature, lots of researchers draw on synthetic datasets. For *Semantic* Pattern Mining, however, the captured data needs to have some semantic expressiveness to mine for meaningful complex patterns.

The situation in Aspect-Oriented Ontology Development is different, and a rich body of ontologies is available for evaluating the approach. Aspect-Oriented Ontology Development is an approach to ontology modularization. The quality of ontology modularizations can be measured using classic modularity metrics. Therefore, the evaluation is straightforward: We measure the modularity quality of existing ontologies, model the same problem using our aspect-oriented approach and compare the module quality with the gold-standard.

Complex Entity Recognition is an approach that combines traditional natural language processing(NLP) and information retrieval(IR) approaches, such as named entity recognition and latent dirichlet allocation, with technologies from semantic web domain and graph mining algorithms in order to mine clusters and correlations of entities from a knowledge graph. However, from a research standpoint this approach is part of the NLP and IR domain, which means that the evaluation methodology and datasets need to be the standard methods and datasets for these research areas. In table 2 we presented some of the most prominent evaluation and trainings datasets from the NLP domain, based on the mention count in the scientific literature. These datasets will form the basis for further scientific evaluation of the proposed approaches and algorithms.

11 The Linking Open Data cloud diagram, last visited April 2016

5 References

[ABK+07] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. *Dbpedia: A nucleus for a web of open data*. Springer, 2007.

[ADGI08] Jagrati Agrawal, Yanlei Diao, Daniel Gyllstrom, and Neil Immerman. Efficient pattern matching over event streams. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 147–160, New York, NY, USA, 2008. ACM.

[ADJ14] Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. How to ask for a favor: A case study on the success of altruistic requests. *CoRR*, abs/1405.3282, 2014.

[AH14] Charu C. Aggarwal and Jiawei Han, editors. *Frequent Pattern Mining*. Springer, 2014.

[AIS93] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, June 1993.

[ALWW09] Charu C. Aggarwal, Yan Li, Jianyong Wang, and Jing Wang. Frequent pattern mining with uncertain data. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 29–38, New York, NY, USA, 2009. ACM.

[AS94] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[Bay98] Roberto J. Bayardo, Jr. Efficiently mining long patterns from databases. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, SIGMOD '98, pages 85–93, New York, NY, USA, 1998. ACM.

[Bay00] Stephen D. Bay. Multivariate discretization of continuous variables for set mining. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, pages 315–319, New York, NY, USA, 2000. ACM.

[BCG01] Doug Burdick, Manuel Calimlim, and Johannes Gehrke. Mafia: A maximal frequent itemset algorithm for transactional databases. In *In ICDE*, pages 443–452, 2001.

[BCM+03] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, New York, NY, USA, 2003.

[BEP+08] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM, 2008.

[BFL98] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.

[BNZ11] Björn Bringmann, Siegfried Nijssen, and Albrecht Zimmermann. Pattern-based classification: A unifying perspective. *CoRR*, abs/1111.6191, 2011.

[CHP05] Shengnan Cong, Jiawei Han, and David Padua. Parallel mining of closed sequential patterns. In *In: Proc. of the 11th ACM SIGKDD Int'l Conf. on Knowledge Discovery in Data Mining*, pages 562–567. ACM Press, 2005.

[CM93] Sharma Chakravarthy and Deepak Mishra. Snoop: An expressive event specification language for active databases, 1993.

[CML11] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1082–1090, New York, NY, USA, 2011. ACM.

[CY05] B. Choi and Z. Yao. Web page classification. In Wesley Chu and Tsau Young Lin, editors, *Foundations and Advances in Data Mining*, volume 180 of *Studies in Fuzziness and Soft Computing*, pages 221–274. Springer, Berlin / Heidelberg, 2005.

[CYKS12] Keunho Choi, Donghee Yoo, Gunwoo Kim, and Yongmoo Suh. A hybrid online-product recommendation system: Combining implicit rating-based collaborative filtering and sequential pattern analysis. *Electronic Commerce Research and Applications*, 11(4):309 – 317, 2012.

[DDPM13] Simon Dooms, Toon De Pessemier, and Luc Martens. Movietweetings: a movie rating dataset collected from twitter. In *Workshop on Crowdsourcing and Human Computation for Recommender Systems, CrowdRec at RecSys 2013*, 2013.

[ED13] Faezeh Ensan and Weichang Du. A semantic metrics suite for evaluating modular ontologies. *Information Systems*, 38(5):745–770, July 2013.

[EHZ06] M. El-Hajj and O. R. Zaiane. Parallel leap: large-scale maximal pattern mining in a distributed environment. In *12th International Conference on Parallel and Distributed Systems - (ICPADS'06)*, volume 1, pages 8 pp.–, 2006.

[FF00] R.E. Filman and D.P. Friedman. Aspect-Oriented Programming Is Quantification and Obliviousness. *Workshop on Advanced Separation of Concerns, OOPSLA*, 2000.

[Fil82] Charles Fillmore. Frame semantics. *Linguistics in the morning calm*, pages 111–137, 1982.

[GH06] Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3), September 2006.

[GJ02] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288, 2002.

[GNPP07] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 330–339, New York, NY, USA, 2007. ACM.

[Gre11] L Green. "common crawl enters a new phase. *Common Crawl blog http://www. commoncrawl. org/common-crawl-enters-a-new-phase*, 2011.

[Gro00] IEEE Architecture Working Group. IEEE standard 1471-2000, Recommended Practice for Architectural Description of Software-Intensive Systems. IEEE, 2000.

[Han05] Jiawei Han. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[HC15] Souleiman Hasan and Edward Curry. Thingsonomy: Tackling variety in internet of things events. *IEEE Internet Computing*, 19(2):10–18, 2015.

[HCXY07] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, 2007.

[HGN00] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. Algorithms for association rule mining - a general survey and comparison. *SIGKDD Explor. Newsl.*, 2(1):58–64, June 2000.

[HLSL00] Farhad Hussain, Huan Liu, Einoshin Suzuki, and Hongjun Lu. *Knowledge Discovery and Data Mining. Current Issues and New Applications: 4th Pacific-Asia Conference, PAKDD 2000 Kyoto, Japan, April 18–20, 2000 Proceedings*, chapter Exception Rule Mining with a Relative Interestingness Measure, pages 86–97. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

[HPY00] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 1–12, New York, NY, USA, 2000. ACM.

[HTP15] Ahmad Hasan, Kia Teymourian, and Adrian Paschke. Spaceroam: An approach to role discovery in collaborative systems. in Proceedings of 1st Workshop on Situation Recognition by Mining Temporal Information SIREMTI2015 at the 7th Conference on Mobile ... MobiCASE2015, to appear., 2015.

[JA07] Ruoming Jin and Gagan Agrawal. *Data Streams: Models and Algorithms*, chapter Frequent Pattern Mining in Data Streams, pages 61–84. Springer US, Boston, MA, 2007.

[KBKW10] Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. The german reference corpus dereko: A primordial sample for linguistic research. In *LREC*, 2010.

[KP03] Paul Kingsbury and Martha Palmer. Propbank: the next level of treebank. In *Proceedings of Treebanks and lexical Theories*, volume 3. Citeseer, 2003.

[Lan95] Ken Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339, 1995.

[Lew97] David D Lewis. Reuters-21578 text categorization test collection, distribution 1.0. *http://www. research. att. com/~ lewis/reuters21578. html*, 1997.

[LFTP15] Alexandra La Fleur, Kia Teymourian, and Adrian Paschke. Complex event extraction from real-time news streams. In *Proceedings of the 11th International Conference on Semantic Systems*, SEMANTICS '15, pages 9–16, New York, NY, USA, 2015. ACM.

[LKLH06] Carson Kai-Sang Leung, Quamrul I. Khan, Zhan Li, and Tariqul Hoque. Cantree: a canonical-order tree for incremental frequent-pattern mining. *Knowledge and Information Systems*, 11(3):287–311, 2006.

[LMFC14] Hoang Thanh Lam, Fabian Mörchen, Dmitriy Fradkin, and Toon Calders. Mining compressing sequential patterns. *Statistical Analysis and Data Mining*, 7(1):34–52, 2014.

[LPW+05] Jinze Liu, S. Paulsen, Wei Wang, A. Nobel, and J. Prins. Mining approximate frequent itemsets from noisy data. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 4 pp.–, Nov 2005.

[Luc12] David C. Luckham. *Event processing for business : organizing the real-time enterprise*. Hoboken, N.J. John Wiley & Sons, 2012.

[LYRL04] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

[mGjW10] Y. m. Guo and Z. j. Wang. A vertical format algorithm for mining frequent item sets. In *Advanced Computer Control (ICACC), 2010 2nd International Conference on*, volume 4, pages 11–13, March 2010.

[ML12] Julian J. McAuley and Jure Leskovec. Image labeling on a network: Using social-network metadata for image classification. *CoRR*, abs/1207.3809, 2012.

[MM02] Gurmeet Singh Manku and Rajeev Motwani. Approximate frequency counts over data streams. In *Proceedings of the 28th International Conference on Very Large Data Bases*, VLDB '02, pages 346–357. VLDB Endowment, 2002.

[MRM+04] Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. The nombank project: An interim report. In *HLT-NAACL 2004 workshop: Frontiers in corpus annotation*, pages 24–31, 2004.

[NZ14] Siegfried Nijssen and Albrecht Zimmermann. *Frequent Pattern Mining*, chapter Constraint-Based Pattern Mining, pages 147–163. Springer International Publishing, Cham, 2014.

[OZ02] Stanley R. M. Oliveira and Osmar R. Zaïane. Privacy preserving frequent itemset mining. In *Proceedings of the IEEE International Conference on Privacy, Security and Data Mining - Volume 14*, CRPIT '14, pages 43–54, Darlinghurst, Australia, Australia, 2002. Australian Computer Society, Inc.

[PBTL99] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In *Proceedings of the 7th International Conference on Database Theory*, ICDT '99, pages 398–416, London, UK, UK, 1999. Springer-Verlag.

[PCT+03] Feng Pan, Gao Cong, Anthony K. H. Tung, Jiong Yang, and Mohammed J. Zaki. Carpenter: Finding closed patterns in long biological datasets. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 637–642, New York, NY, USA, 2003. ACM.

[PHM00] Jian Pei, Jiawei Han, and Runying Mao. Closet: An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 21–30, 2000.

[PS09] Christine Parent and Stefano Spaccapietra. An Overview of Modularity. In Heiner Stuckenschmidt, Christine Parent, and Stefano Spaccapietra, editors, *Modular Ontologies*, number 5445 in Lecture Notes in Computer Science, pages 5–23. Springer Berlin Heidelberg, January 2009. DOI: 10.1007/978-3-642-01907-4.

[PST+14] Adrian Paschke, Ralph Schäfermeier, Kia Teymourian, Alexandru Todor, and Ahmad Hassan. Corporate Smart Content: Requirements and Use Cases. Report I on the sub-project Smart Content Enrichment. Technical Report TR-B-14-02, Freie Universität Berlin, 2014.

[PTCX04] Feng Pan, A. K. H. Tung, Gao Cong, and Xin Xu. Cobbler: combining column and row enumeration for closed pattern discovery. In *Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on*, pages 21–30, June 2004.

[PZC+03] Jian Pei, Xiaoling Zhang, Moonjung Cho, Haixun Wang, and P. S. Yu. Maple: a fast algorithm for maximal pattern-based clustering. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 259–266, Nov 2003.

[RA15] R. A. Rossi and N. K. Ahmed. Role discovery in networks. *IEEE Transactions on Knowledge and Data Engineering*, 27(4):1112–1131, April 2015.

[Res95] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[SB12] Mohammad Karim Sohrabi and Ahmad Abdollahzadeh Barforoush. Efficient colossal pattern mining in high dimensional datasets. *Knowledge-Based Systems*, 33:41 – 52, 2012.

[Sch91] Klaus Schild. A correspondence theory for terminological logics: Preliminary report. In John Mylopoulos and Raymond Reiter, editors, *Proceedings of the 12th International Joint Conference on Artificial Intelligence. Sydney, Australia, August 24-30, 1991*, pages 466–471. Morgan Kaufmann, 1991.

[SL13] Thabet Slimani and Amor Lazzez. Sequential mining: Patterns and algorithms analysis. *CoRR*, abs/1311.0350, 2013.

[SMMP09] Nicholas Poul Schultz-Møller, Matteo Migliavacca, and Peter Pietzuch. Distributed complex event processing with query rewriting. In *Proceedings of the Third ACM International Conference on Distributed Event-Based Systems*, DEBS '09, pages 4:1–4:12, New York, NY, USA, 2009. ACM.

[SNV07] L. Szathmary, A. Napoli, and P. Valtchev. Towards rare itemset mining. In *19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007)*, volume 1, pages 305–312, Oct 2007.

[Ste05] Friedrich Steimann. Domain Models Are Aspect Free. In Lionel Briand and Clay Williams, editors, *Model Driven Engineering Languages and Systems*, number 3713 in Lecture Notes in Computer Science, pages 171–185. Springer Berlin Heidelberg, January 2005.

[TAJL08] Syed Khairuzzaman Tanbeer, Chowdhury Farhan Ahmed, Byeong-Soo Jeong, and Young-Koo Lee. *Advances in Knowledge Discovery and Data Mining: 12th Pacific-Asia Conference, PAKDD 2008 Osaka, Japan, May 20-23, 2008 Proceedings*, chapter CP-Tree: A Tree Structure for Single-Pass Frequent Pattern Mining, pages 1022–1027. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[TBA+14] R. Tönjes, P. Barnaghi, M. Ali, A. Mileo, M. Hauswirth, F. Ganz, S. Ganea, B. Kjærgaard, D. Kuemper, S. Nechifor, D. Puiu, A. Sheth, V. Tsiatsis, and L. Vestergaard. Real time iot stream processing and large-scale data analytics for smart city applications, 2014. Poster presented at European Conference on Networks and Communications.

[VK14] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[Wel10] Chris Welty. Context slices: representing contexts in owl. In *Workshop on Ontology Patterns*, page 59, 2010.

[WWZ+05] Wei Wang, Chen Wang, Yongtai Zhu, Baile Shi, Jian Pei, Xifeng Yan, and Jiawei Han. Graphminer: A structural pattern-mining system for large disk-based graph databases and its applications. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, pages 879–881, New York, NY, USA, 2005. ACM.

[XHYC05] Dong Xin, Jiawei Han, Xifeng Yan, and Hong Cheng. Mining compressed frequent-pattern sets. In *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, pages 709–720. VLDB Endowment, 2005.

[YC05] Chung-Ching Yu and Yen-Liang Chen. Mining sequential patterns from multidimensional sequence data. *IEEE Transactions on Knowledge and Data Engineering*, 17(1):136–140, Jan 2005.

[YCL+06] Jeffrey Xu Yu, Zhihong Chong, Hongjun Lu, Zhenjie Zhang, and Aoying Zhou. A false negative approach to mining frequent itemsets from high speed transactional data streams. *Information Sciences*, 176(14):1986 – 2015, 2006. Streaming Data Mining.

[ZH] Mohammed J. Zaki and Ching-Jui Hsiao. *CHARM: An Efficient Algorithm for Closed Itemset Mining*, chapter 27, pages 457–473.

[ZPOL97] Mohammed J Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, and Wei Li. New algorithms for fast discovery of association rules. Technical report, Rochester, NY, USA, 1997.

6 Appendix

Table 6: Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
Bao	2006	Wine/Food	SHOIN(D) (OWL 2 DL)	77	657
		http://danielsmith.eu/resources/facet/#	ALE(D) (OWL 2 DL)	7	37
Spyns	2002	BibliOntology	ALUHIN(D) (OWL 2 DL)	69	299
Pathak	2009	Juvenile Rheumatoid Arthritis (JRA)			
Jiménez-Ruiz	2008	Juvenile Rheumatoid Arthritis (JRA)			
Pathak	2009	NCI			
Pathak	2009	SNOMED CT			
Pathak	2009	OpenGALEN8 CRM	AL (OWL 2)	0	0
Pathak	2009	OpenGALEN8 DD 104 ChapterVII Obstetrics	ALEI(D) (OWL 2)	1856	2167
Pathak	2009	OpenGALEN8 DD 111 ChapterIII Respiratory	ALEI (OWL 2)	1001	1272
Pathak	2009	OpenGALEN8 DD 113 ChapterX Musculoskeletal	ALEI(D) (OWL 2)	1669	2119
Pathak	2009	OpenGALEN8 DD 131 ChapterVIII Oncology	ALEI (OWL 2)	1884	2225
Pathak	2009	OpenGALEN8 DD 136 ChapterI Gastrointestinal	ALEI (OWL 2)	1889	2263

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
Pathak	2009	OpenGALEN8 DD 144 ChapterIX Nutrition	ALEI(D) (OWL 2)	1547	1697
Pathak	2009	OpenGALEN8 DD 161 ChapterXIII Skin	ALEI(D) (OWL 2)	1627	1971
Pathak	2009	OpenGALEN8 DD 164 ChapterVI Endocrine	ALEI(D) (OWL 2)	2906	3753
Pathak	2009	OpenGALEN8 DD 226 ChapterV Infection	ALEI (OWL 2)	3404	5288
Pathak	2009	OpenGALEN8 DD 2383 VMP A	ALEI (OWL 2)	3417	2249
Pathak	2009	OpenGALEN8 DD 297 ChapterIV CNS	ALEI(D) (OWL 2)	4527	9628
Pathak	2009	OpenGALEN8 DD 2 Chapters	ALEI (OWL 2)	15	9
Pathak	2009	OpenGALEN8 DD 346 Reinstated	ALEI (OWL 2)	4550	5126
Pathak	2009	OpenGALEN8 DD 38 ChapterXIV Immunology	ALEI (OWL 2)	214	169
Pathak	2009	OpenGALEN8 DD 3914 VMP B	ALEI (OWL 2)	5510	7411
Pathak	2009	OpenGALEN8 DD 409 Abstractions	ALEI (OWL 2)	1149	985
Pathak	2009	OpenGALEN8 DD 4319 Interactions	ALEI(D) (OWL 2)	3746	3239
Pathak	2009	OpenGALEN8 DD 474 ChapterII Cardiovascular	ALEI(D) (OWL 2)	4936	7628
Pathak	2009	OpenGALEN8 DD 476 BNFInteractions	ALEI(D) (OWL 2)	3695	3165
Pathak	2009	OpenGALEN8 DD 490 NewInteractions	ALEI(D) (OWL 2)	3701	3187
Pathak	2009	OpenGALEN8 DD 55 DrugsWithNoDissection	ALEI (OWL 2)	145	70
Pathak	2009	OpenGALEN8 DD 57 AdditionalInteractions	ALEI(D) (OWL 2)	114	76

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
Pathak	2009	OpenGALEN8 DD 635 Formulations	ALEI (OWL 2)	963	744
Pathak	2009	OpenGALEN8 DD 76 ChapterXV Anaesthesia	ALEI(D) (OWL 2)	850	851
Pathak	2009	OpenGALEN8 DD 8048 Components	ALEI(D) (OWL 2)	15643	13361
Pathak	2009	OpenGALEN8 DD 84 ChapterXII ENT	ALEI (OWL 2)	861	838
Pathak	2009	OpenGALEN8 DD 91 ChapterXI Eye	ALEI (OWL 2)	1168	1175
Pathak	2009	OpenGALEN8 DissectionsDisease	ALEI(D) (OWL 2)	12210	10656
Pathak	2009	OpenGALEN8 DissectionsDrugOntology	ALI (OWL 2)	8	7
Pathak	2009	OpenGALEN8 DissectionsModel	ALEHIF (OWL 2)	393	201
Pathak	2009	OpenGALEN8 DissectionsSurgicalProcedure	ALI (OWL 2)	8	7
Pathak	2009	OpenGALEN8 DS 1064 Sensory	ALEI(D) (OWL 2)	4922	4358
Pathak	2009	OpenGALEN8 DS 118 Lymphoreticular	ALEI (OWL 2)	403	321
Pathak	2009	OpenGALEN8 DS 1507 Vascular	ALEI(D) (OWL 2)	7403	7037
Pathak	2009	OpenGALEN8 DS 205 Endocrine	ALEI(D) (OWL 2)	705	604
Pathak	2009	OpenGALEN8 DS 209 Respiratory	ALEI(D) (OWL 2)	806	685
Pathak	2009	OpenGALEN8 DS 2157 Genitourinary	ALEI(D) (OWL 2)	8748	8047
Pathak	2009	OpenGALEN8 DS 219 MinorSurgery	ALEI(D) (OWL 2)	1559	1335
Pathak	2009	OpenGALEN8 DS 224 Therapeutic	ALEI(D) (OWL 2)	1156	898

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
Pathak	2009	OpenGALEN8 DS 307 Abstractions	ALEI(D) (OWL 2)	1112	892
Pathak	2009	OpenGALEN8 DS 3227 Cardiothoracic	ALEI(D) (OWL 2)	10621	9878
Pathak	2009	OpenGALEN8 DS 326 Ophthalmology	ALEI (OWL 2)	1928	1683
Pathak	2009	OpenGALEN8 DS 4076 Musculoskeletal	ALEI(D) (OWL 2)	16170	15922
Pathak	2009	OpenGALEN8 DS 425 Paediatric	ALEI(D) (OWL 2)	3155	2768
Pathak	2009	OpenGALEN8 DS 528 Neurosurgery	ALEI(D) (OWL 2)	2785	2483
Pathak	2009	OpenGALEN8 DS 54 Breast	ALEI(D) (OWL 2)	284	218
Pathak	2009	OpenGALEN8 DS 65 Demo	ALEI(D) (OWL 2)	257	199
Pathak	2009	OpenGALEN8 DS 711 Orodental	ALEI(D) (OWL 2)	3633	3284
Pathak	2009	OpenGALEN8 DS 755 SkinPlastic	ALEI(D) (OWL 2)	3847	3518
Pathak	2009	OpenGALEN8 DS 85 Diagnostic	ALEI(D) (OWL 2)	531	418
Pathak	2009	OpenGALEN8 DS 935 Digestive	ALEI(D) (OWL 2)	3826	3423
Pathak	2009	OpenGALEN8 FoundationModel	ALI (OWL 2)	8	7
Pathak	2009	OpenGALEN8 FoundationModel ClinicalSituationModel	ALI (OWL 2)	9	7
Pathak	2009	OpenGALEN8 FoundationModel DetailedMedicalCategorySpace	ALI (OWL 2)	2988	2972
Pathak	2009	OpenGALEN8 FoundationModel GRAILSanctions	ALEI (OWL 2)	1089	1103

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
Pathak	2009	OpenGALEN8 FoundationModel NamedDefinedEntities	ALEI(D) (OWL 2)	335	188
Pathak	2009	OpenGALEN8 FoundationModel TopClassesForMedical-Domain	ALI (OWL 2)	898	926
Pathak	2009	OpenGALEN8 FULL	AL (OWL 2)	0	0
Pathak	2009	OpenGALEN8 GenericModel	ALI (OWL 2)	28	26
Pathak	2009	OpenGALEN8 GenericModel GRAILSanctions	ALEI(D) (OWL 2)	42	112
Pathak	2009	OpenGALEN8 GenericModel PropertyChainAxioms	ALRI (OWL 2)	8	163
Pathak	2009	OpenGALEN8 GenericModel PropertyTypes	ALHIF (OWL 2)	8	1724
Pathak	2009	OpenGALEN8 GenericModel TopOntologyOfClasses	ALI (OWL 2)	560	581
Pathak	2009	OpenGALEN8 MedicalExtensions	ALI (OWL 2)	8	7
Pathak	2009	OpenGALEN8 MedicalExtensions Abstractions	ALEI (OWL 2)	75	51
Pathak	2009	OpenGALEN8 MedicalExtensions BasicMedicalScience	ALI (OWL 2)	8	7
Pathak	2009	OpenGALEN8 MedicalExtensions Biochemistry	ALEI (OWL 2)	667	582
Pathak	2009	OpenGALEN8 MedicalExtensions Chemistry	ALEI (OWL 2)	128	95
Pathak	2009	OpenGALEN8 MedicalExtensions ClinicalPragmatics	ALEHI (OWL 2)	114	88
Pathak	2009	OpenGALEN8 MedicalExtensions Devices	ALEI (OWL 2)	1107	855
Pathak	2009	OpenGALEN8 MedicalExtensions Drugs	ALEHIF(D) (OWL 2)	4258	5042

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
Pathak	2009	OpenGALEN8 MedicalExtensions Genetics	ALEI (OWL 2)	51	45
Pathak	2009	OpenGALEN8 MedicalExtensions HealthcareOrganisation	ALEI (OWL 2)	85	47
Pathak	2009	OpenGALEN8 MedicalExtensions HumanAnatomy	ALEHIF (OWL 2)	7859	17318
Pathak	2009	OpenGALEN8 MedicalExtensions Investigations	ALEI (OWL 2)	1176	876
Pathak	2009	OpenGALEN8 MedicalExtensions Microbiology	ALEI (OWL 2)	965	1013
Pathak	2009	OpenGALEN8 MedicalExtensions Pathology	ALEI(D) (OWL 2)	7378	7159
Pathak	2009	OpenGALEN8 MedicalExtensions Physiology	ALEI (OWL 2)	2129	2256
Pathak	2009	OpenGALEN8 MedicalExtensions Psychiatry	ALEHIF (OWL 2)	231	208
Pathak	2009	OpenGALEN8 MedicalExtensions SignsSymptoms	ALEI (OWL 2)	1431	947
Pathak	2009	OpenGALEN8 MedicalExtensions SocialAndEnvironment	ALEHIF (OWL 2)	425	319
Pathak	2009	OpenGALEN8 MedicalExtensions SurgicalProcedures	ALEI (OWL 2)	1033	673
-	-	Amphibian gross anatomy	ALE (OWL 2 EL)	1603	2672
-	-	Anatomical Entity Ontology	ALE (OWL 2 EL)	250	366
-	-	Adverse Event Reporting Ontology	ALCHOIQ (OWL 2)	202	1101
-	-	Ascomycete phenotype ontology	AL (OWL 2 EL)	342	306

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
-	-	Amphibian taxonomy	ALE (OWL 2 EL)	6135	12163
-	-	Biological Collections Ontology	SROIF(D) (OWL 2)	127	304
-	-	Basic Formal Ontology	ALC (OWL 2 EL)	35	52
-	-	Bilateria anatomy	ALEHI+ (OWL 2 DL)	114	139
-	-	Biological Spatial Ontology	ALERI+ (OWL 2 DL)	146	473
-	-	BRENDA tissue / enzyme source	ALE (OWL 2 EL)	5809	6884
-	-	Common Anatomy Reference Ontology	ALE+ (OWL 2 EL)	48	50
-	-	Comparative Data Analysis Ontology	SROIQ(D) (OWL 2)	132	421
-	-	Cephalopod Ontology	SRI (OWL 2 DL)	410	604
-	-	Chemical Entities of Biological Interest	ALE+ (OWL 2 EL)	61943	163061
-	2016	Chemical Information Ontology	SRIF(D) (OWL 2)	329	467
-	2016	Chemical Methods Ontology	ALCH (OWL 2 EL)	2965	3417

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
Diehl	2016	Cell Ontology	ALCR (OWL 2 DL)	3323	8106
Diehl	2016	Cell Line Ontology	ALEO (OWL 2)	38994	51292
-	-	Clinical measurement ontology	ALE+ (OWL 2 EL)	2680	3297
-	-	Ctenophore Ontology	ALE (OWL 2)	116	125
-	-	Cardiovascular Disease Ontology	ALCR (OWL 2)	486	580
-	-	Dictyostelium discoideum anatomy	ALE+ (OWL 2 EL)	131 347	-
-	-	Dictyostelium discoideum phenotype	AL (OWL 2 EL)	972	1109
-	-	Drug Interaction and Evidence Ontology	SRIQ (OWL 2)	157	254
-	-	Human Disease Ontology	AL (OWL 2 EL)	9201	7040
-	-	Drosophila Phenotype Ontology	SRIF(D) (OWL 2 DL)	505	1272
-	-	Evidence ontology	ALCI (OWL 2 DL)	683	1397
-	-	Human developmental anatomy, timed version	ALE (OWL 2 EL)	8340	8339
-	-	Human developmental anatomy, abstract version	ALE (OWL 2 EL)	2314	2335
-	-	Human developmental anatomy, abstract	ALE+ (OWL 2 EL)	2734	13369

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
-	-	Mouse gross anatomy and development, timed	ALE+ (OWL 2 EL)	19444	21723
-	-	Mouse gross anatomy and development, timed	ALE+ (OWL 2 EL)	6239	21772
Envo	2016	Environment Ontology	SRI (OWL 2)	6131	9882
Jaiswal	2015	Plant Environment Ontology	AL (OWL 2 EL)	561	557
-	-	eagle-i resource ontology	SHOIF(D) (OWL 2)	4061	4893
Mattingly	2015	Exposure ontology	ALER+ (OWL 2 EL)	81	101
-	-	Fungal gross anatomy	ALEI+ (OWL 2 DL)	90	115
-	-	Biological imaging methods	S (OWL 2 EL)	624	596
-	-	Drosophila gross anatomy	SRI (OWL 2 DL)	9642	31346
-	-	FlyBase Controlled Vocabulary	SRIF(D) (OWL 2 DL) 1542	2643	-
-	-	Drosophila development	SHI (OWL 2 DL)	234	660
-	2013	Fly taxonomy	AL (OWL 2 EL)	6599	6585
-	-	Physico-chemical methods and properties	ALE (OWL 2 EL)	1163	1684

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
SquireS	2015	Influenza Ontology	ALCI (OWL 2)	213	204
Mejino	2016	Foundational Model of Anatomy (subset)	ALEH+ (OWL 2 EL)	78977	121712
-	-	Fission Yeast Phenotype Ontology	SH (OWL 2 EL)	9335	30756
AshburneR	2014	Gazetteer	ALRI+ (OWL 2)	10685	898145
-	-	Geographical Entity Ontology	ALCHQ (OWL 2)	35	498
-	2016	Gene Ontology	SRI (OWL 2 DL)	44338	103077
-	-	Cereal Plant Gross Anatomy	AL (OWL 2 EL)	0	0
-	-	Hymenoptera Anatomy Ontology	SR (OWL 2 EL)	2349	9996
-	-	Homology Ontology	ALC (OWL 2 EL)	66	83
-	-	human phenotype ontology	ALCR (OWL 2)	13733	20985
-	-	Human Developmental Stages	ALEH+ (OWL 2 EL)	221	647
-	-	Information Artifact Ontology	SROIN(D) (OWL 2 DL)	180	383
-	-	Infectious disease	AL (OWL 2)	0	0
Topalis	2015	Malaria Ontology	ALERI+ (OWL 2 DL)	2598	3468

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
-	-	Event (INOH pathway ontology)	ALEH+ (OWL 2 EL)	3828	7125
-	-	Molecule role (INOH Protein name/family name ontology)	ALE+ (OWL 2 EL)	9217	9623
-	-	Loggerhead nesting	ALE (OWL 2 EL)	308	347
-	-	Mouse adult gross anatomy	ALE+ (OWL 2 EL)	3229	4103
-	-	Minimal anatomical terminology	ALE (OWL 2 EL)	461 504	-
-	-	Mental Functioning Ontology	ALE (OWL 2)	30	21
HastinGs	2016	Emotion Ontology	ALEI (OWL 2)	227	289
-	-	MFO Mental Disease Ontology	AL (OWL 2)	13	6
-	-	Molecular interactions	ALE+ (OWL 2 EL)	1424	1406
-	-	MIAPA Ontology	ALEHI(D) (OWL 2)	54	73
-	-	prokaryotic phenotypic and metabolic characters	SROIQ(D) (OWL 2)	3991	5671
DritsoU	2015	microRNA Ontology	ALEI (OWL 2 QL)	676	764

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
-	-	Mosquito insecticide resistance	ALE+ (OWL 2 EL)	4408	4457
-	-	Measurement method ontology	AL (OWL 2 EL)	566	638
-	-	Mouse Developmental Stages	ALEH+ (OWL 2 EL)	111	299
Natale	2016	Protein modification	ALE+ (OWL 2 EL)	2001	3581
-	-	Molecular Process Ontology	ALCH (OWL 2 DL)	3635	3809
-	-	Mammalian phenotype	ALC (OWL 2)	15369	22281
-	-	Mouse pathology	ALE+ (OWL 2 EL)	888	941
-	-	MHC Restriction Ontology	ALEH (OWL 2 EL)	1768	3098
-	-	Mass spectrometry ontology	ALE+ (OWL 2)	1825	2220
-	-	Neuro Behavior Ontology	ALC (OWL 2)	1083	1402
-	-	NCBI organismal classification	AL (OWL 2 EL)	1410459	1410457
CharlEt	2011	Ontology of Adverse Events	ALCH (OWL 2)	3379	4388
CharlEt	2011	Ontology of Biological Attributes	ALERI+ (OWL 2 DL)	4259	13828

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
ZhenG StoeckErt BrochhauSen	2015	Ontology for Biomedical Investigations	SROIQ(D) (OWL 2)	2933	6134
-	-	The Ontology of Genes and Genomes	SRIQ (OWL 2)	69689	70103
ZhenG StoeckErt BrochhauSen	2015	Ontology for genetic interval	ALCHIQ(D) (OWL 2)	156	372
-	-	Ontology for General Medical Science	AL (OWL 2 DL)	86	79
-	-	Ontology of Genetic Susceptibility Factor	ALCHIQ(D) (OWL 2 DL)	95	235
-	-	Medaka Developmental Stages	ALEH+ (OWL 2 EL)	47	91
-	-	Ontologized MIABIS	ALCRIQ (OWL 2)	193	255
ZhenG StoeckErt BrochhauSen	2015	Ontology for MIRNA Target	ALEHI+ (OWL 2 DL)	2700	2746
CharlEt	2011	Ontology of Microbial Phenotypes	ALER+ (OWL 2 EL)	1109	1495
-	-	Ontology of Medically Related Social Entities	ALCRIQ (OWL 2 DL)	196	249
-	-	Ontology of Vaccine Adverse Events	ALCH(D) (OWL 2)	794	1766
Jaiswal	2015	Plant Anatomy Ontology	AL (OWL 2 EL)	0	0

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
-	-	Phenotypic quality	SH (OWL 2 EL)	2502	2347
-	-	Platynereis Developmental Stages	ALEH+ (OWL 2 EL)	21	42
-	-	Plant Growth and Development Stage	AL (OWL 2 EL)	0	0
Jaiswal	2015	Plant Ontology	SRI (OWL 2 DL)	1728	2860
ThacKer	2015	Porifera Ontology	ALCRQ (OWL 2 DL)	674	875
-	-	Proteomics data and process provenance	SHOIN(D) (OWL 2)	399	722
Petri	2016	Pathway ontology	ALE (OWL 2 EL)	2537	3134
Chebi	2011	Physico-chemical process	ALE (OWL 2 EL)	552	730
BatcheLor	2015	RNA ontology	SRIQ (OWL 2 DL)	692	1864
-	-	Relations Ontology	SRIF (OWL 2)	51	792
-	-	Rat Strain Ontology	ALE (OWL 2 EL)	4082	5731
-	-	Name Reaction Ontology	ALCH (OWL 2 DL)	759	1299
-	-	Systems Biology	AL (OWL 2 EL)	609	639
-	-	Sample processing and separation techniques	ALE (OWL 2)	228	261

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
-	-	Social Insect Behavior Ontology	ALE+ (OWL 2 EL)	525	1106
-	-	Suggested Ontology for Pharmacogenomics	ALCHIN(D) (OWL 2)	108	282
Ramirez	2014	Spider Ontology	ALE+ (OWL 2 EL)	661	840
-	-	The Statistical Methods Ontology	SROIQ(D) (OWL 2 DL)	609	1764
SchriMl	2013	Symptom Ontology	AL (OWL 2 EL)	936	840
-	-	Tick gross anatomy	ALE+ (OWL 2 EL)	628	948
-	-	Teleost Anatomy Ontology	ALERI+ (OWL 2 DL)	3372	5190
-	-	Taxonomic rank vocabulary	AL (OWL 2 EL)	59	58
-	-	Mosquito gross anatomy	ALE+ (OWL 2 EL)	1864	2733
Jaiswal	2015	Plant Trait Ontology	ALE (OWL 2 DL)	1721	3285
-	-	Pathogen Transmission Ontology	AL (OWL 2 EL)	28	24
-	-	Teleost taxonomy	AL (OWL 2 EL)	38704	38639

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
-	-	Uberon multi-species anatomy ontology	SRIQ (OWL 2)	14335	45588
Arabandi	2010	Units of measurement	ALE (OWL 2 EL)	379	389
-	-	Combined phenotype ontology	AL (OWL 2 EL)	28925	41179
VihineN	2015	Variation Ontology	ALER+ (OWL 2 EL)	390	405
-	-	Vertebrate Homologous Ontology Group Ontology	ALE+ (OWL 2 EL)	1185	1688
He	2016	Vaccine Ontology	ALCHOQ (OWL 2)	4303	9647
-	-	Vertebrate Skeletal Anatomy Ontology-	ALERI+ (OWL 2 DL)	314	457
-	-	Vertebrate trait	AL (OWL 2 EL)	3299	3757
-	-	Vertebrate Taxonomy Ontology	AL (OWL 2 EL)	107134	106940
-	-	C. elegans gross anatomy	S (OWL 2 DL)	7236	12414
-	-	C. elegans development	ALEH+ (OWL 2 EL)	707	1394
-	-	C. elegans phenotype	AL (OWL 2 EL)	2396	2718
-	-	Xenopus anatomy and development	ALE+ (OWL 2 EL)	1521	5887

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
-	-	Experimental condition ontology	ALE+ (OWL 2 EL)	494	595
-	-	Yeast phenotypes	AL (OWL 2 EL)	300	266
-	-	bcgo merged inferred	SROIN(D) (OWL 2)	2270	4514
-	-	dron-full	AL (OWL 2 DL)	0	0
-	-	epidemiology ontology	ALC (OWL 2 DL)	191	296
-	-	zfa	SRI (OWL 2 DL)	3053	11624
-	-	feed	ALCHIQ (OWL 2)	214	353
-	-	zfs	ALEH+ (OWL 2 EL)	54	148
-	-	ico merged	SHOIN(D) (OWL 2)	375	705
-	-	kisao full	ALC (OWL 2 DL)	42	50
-	-	mamo-xml	ALCR (OWL 2 DL)	100	164
-	-	ncro-all-in-one	SROI (OWL 2 DL)	3078	5471
-	-	obcs merged inferred	SROIQ(D) (OWL 2 DL)	779	1422

Table 6: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
-	-	obib merged inferred	SROIQ(D) (OWL 2 DL)	522	1194
-	-	opl inferred	SHOIF (OWL 2)	361	886
-	-	pco merged inferred	SROIQ (OWL 2)	1546	2537
-	-	pro	SQ (OWL 2)	296493	540576
-	-	so-xp	SHI (OWL 2 DL)	2301	3021
-	-	swo merged	SHOIQ(D) (OWL 2)	4068	7683

Table 7: Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
Shvaiko	2013	Mouse adult gross anatomy	ALE+	3229	4103
Shvaiko	2013	NCI Thesaurus	SH(D)	117257	164804
Noy	2004	Suggested Upper Merged Ontology (SUMO)	SOIF	4558	175208
Noy	2004	DOLCE : a Descriptive Ontology for Linguistic and Cognitive Engineering	SHOIN(D)	245	1722
Seddiqui	2015	Abox	AL	0	0

Table 7: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
Seddiqui	2015	Tbox	AL	0	0
Xiang	2015	OAEI Benchmark Biblio 2012 (1 reference ontology, 94 target ontologies)			
Xiang	2015	OAEI Benchmark Biblio 2013 (1 reference ontology, 93 target ontologies)			
Xiang	2015	OAEI Standard Benchmark 2012 (1 reference ontology, 109 target ontologies)			
Dimou	2015	DBLP			
Dimou	2015	Contact Details of Flemish Local Governments Dataset (CDFLG)	AL	7	106
Dimou	2015	iLastic			
Dimou	2015	CEUR-WS from ESWC2015 Semantic Publishing Challenge (SPC)			
Dimou	2015	DBpedia	ALCHF(D)	1165	7111
Zarembo	2015	Land Administration Domain Model (LADM)			
Arnold	2015	DBpedia	ALCHF(D)	1165	7111
Arnold	2015	YAGO	AL	172	101
Arnold	2015	BabelNet			
Sherif	2015	Drugbank	AL	76	184

Table 7: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
Sherif	2015	DBpedia	ALCHF(D)	1165	7111
Sherif	2015	Jamendo			
Jimenez-Ruiz	2015	BootOX			
Nosner	2015	DBpedia	ALCHF(D)	1165	7111
Nosner	2015	Freebase			
Holub	2015	DBpedia	ALCHF(D)	1165	7111
Holub	2015	YAGO	AL	172	101
Pirro	2015	DBpedia	ALCHF(D)	1165	7111
Pirro	2015	Linked Movie Database (LinkedMDB)			
Pirro	2015	Google Knowledge Graph (KG)			
Laurini	2015	Geonames	ALEO	150	336
Laurini	2015	GeoSPARQL	SHIF(D)	78	191
Ramar	2015	OAEI Benchmark Biblio 2007			
Saveta	2015	OAEI Benchmark IIMB 2009			
Saveta	2015	ONTOBI			
Saveta	2015	SWING			
Anam	2015	AGROVOC	AL	15960	16382
Anam	2015	EUROVOC	SRIN(D)	20	224

Table 7: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
Hu	2015	Semantic Web for Earth and Environmental Terminology (SWEET)	SHOIN(D)	4550	10216
Dragoni	2015	AGROVOC	AL	15960	16382
Dragoni	2015	EUROVOC	SRIN(D)	20	224
Jiang	2015	Nba-os			
Jiang	2015	Census – income			
Jiang	2015	OntoFarm – cmt	ALCIN(D)	30	226
Jiang	2015	OntoFarm – conference	ALCOI(D)	62	247
Jiang	2015	OntoFarm – edas	ALCOIN(D)	104	739
Jiang	2015	OntoFarm – ekaw	SHIN	74	233
Son	2015	OAEI 2011 Benchmark	ALUHON(D)	37	562
Xue	2015	OAEI 2011 Benchmark	ALUHON(D)	37	562
Pinkel	2015	OntoFarm – cmt	ALCIN(D)	30	226
Pinkel	2015	OntoFarm – conference	ALCOI(D)	62	247
Pinkel	2015	OntoFarm – SIGKDD	ALEI(D)	50	116
Anam	2015	XDR Schema: EXCEL			
Anam	2015	XDR Schema: CIDX			
Anam	2015	XDR Schema: NORIS			

Table 7: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
Anam	2015	XDR Schema: PARAGON			
Guang Zheng	2015	IOOS Platform Vocabulary	AL(D)	1	183
Guang Zheng	2015	IOOS Parameter Vocabulary	AL(D)	1	743
Guang Zheng	2015	MMI Platform Ontology	SH	163	311
Guang Zheng	2015	Climate and Forecast (CF) standard names parameter vocabulary	AL(D)	2	5342
Guang Zheng	2015	DRDC Atlantic NADAS Parameter Codes	AL(D)	1	125
Kejriwal	2015	OAEI 2010 Persons 1	ALRF(D)	10	40
Kejriwal	2015	OAEI 2010 Persons 2	ALRF(D)	10	40
Kejriwal	2015	OAEI 2010 Restaurants	ALR(D)	5	18
Kejriwal	2015	ACM-DBLP			
del Carmen Legaz-Garcia	2015	openEHR (electronic healthcare record)	ALUOIN(D)	144	764
del Carmen Legaz-Garcia	2015	CEM (Clinical Element Model)	ALCHIQ(D)	28	103
Wang	2015	DBLP			
Wang	2015	Query Log			
Fany	2014	Freebase			
Diallo	2014	NCI Thesaurus	SH(D)	117257	164804

Table 7: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
Diallo	2014	Foundational Model of Anatomy (FMA)	ALEH+	78977	121712
Diallo	2014	Thesaurus for the Social Sciences (TheSoz)	ALC(D)	8394	39416
Dutta	2014	DBpedia	ALCHF(D)	1165	7111
Dutta	2014	YAGO	AL	172	101
Dutta	2014	Freebase			
Zeng	2014	DBLP			
Hu	2014	DBpedia	ALCHF(D)	1165	7111
Hu	2014	DBtune			
Hu	2014	LinkedMDB			
Locoro	2014	AGROVOC	AL	15960	16382
Locoro	2014	NAL from US DoA			
Solimando	2014	Norwegian Petroleum Directorate (NPD) FactPages Vocabulary	AL	46	26
Sabbah	2014	AGROVOC	AL	15960	16382
Sabbah	2014	Chinese Agricultural Thesaurus (CAT)			
Sleeman	2014	DBpedia	ALCHF(D)	1165	7111
Sleeman	2014	Freebase			
Sleeman	2014	Arnetminer			
Xu	2014	DBpedia	ALCHF(D)	1165	7111

Table 7: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
Xu	2014	Geonames	ALEO	150	336
Xu	2014	LinkedMDB			
Jauro	2014	OAEI Dataset			
Pinkel	2014	Music Ontology			
Gong	2014	DBpedia	ALCHF(D)	1165	7111
Gong	2014	Freebase			
Gong	2014	Geonames	ALEO	150	336
Cochez	2014	Foundational Model of Anatomy (FMA)	ALEH+	78977	121712
Cochez	2014	NCI Thesaurus	SH(D)	117257	164804
Cochez	2014	Systematized Nomenclature of Human and Veterinary Medicine (SNOMED)			
Chiang	2014	European Patent data			
Chiang	2014	DBLP			
Cruz	2014	OAEI Benchmarks 2010			
Raunich	2014	Mouse adult gross anatomy	ALE+	3229	4103
Raunich	2014	NCI Thesaurus	SH(D)	117257	164804
Raada	2014	OAEI 2011 Benchmark	ALUHON(D)	37	562
Faria	2014	Bone Dysplasia Ontology (BDO)	SHIF(D)	13817	45956

Table 7: (Cont.) Ontologies relevant in the field of modularization

Author	Year	Ontology Name	Expressivity	Class count	Logical axiom count
Faria	2014	Cell Culture Ontology (CCONT)	SRI	18103	30808
Faria	2014	Experimental Factor Ontology (EFO)	ALHI+	17263	24397
Faria	2014	Mouse adult gross anatomy	ALE+	3229	4103
Faria	2014	Cardiac Electrophysiology Ontology (EP)	SHF(D)	82066	172715
Faria	2014	Foundational Model of Anatomy (FMA)	ALEH+	78977	121712
Faria	2014	NCI Thesaurus	SH(D)	117257	164804
Faria	2014	Sleep Domain Ontology (SDO)	SHOIQ(D)	1382	2857
Faria	2014	Single-Nucleotide Polymorphism (SNP)	SHIN(D)	2286	10914
Faria	2014	Sequence Types and Features Ontology (SO)	SHI	2301	3021
Faria	2014	Uber Anatomy Ontology (UBERON)	SRIQ	14323	45323
Faria	2014	Teleost Anatomy Ontology (TAO)	ALERI+	3372	5190
Faria	2014	Zebrafish Anatomy and Development Ontology (ZFA)	SRI	3053	11624
Ngomo	2014	OAEI Benchmarks 2010			