K-CAP2017 Satellites - Workshops and Tutorials

Carlos Badenes
Universidad Politecnica de Madrid
cbadenes@fi.upm.es

Ronald Denaux
Expert System
rrdenaux@expertsystem.com

Martine De Vos
Netherlands eScience Centre
m.devos@esciencecenter.nl

Daniel Garijo
University of Southern California
dgarijo@isi.edu

Jose Manuel Gomez-Perez
Expert System
jmgomez@expertsystem.com

Agnieszka Lawrynowicz
Poznan University
agnieszka.lawrynowicz@cs.put.poznan.pl

Pasquale Lisena
EURECOM
pasquale.lisena@eurecom.fr

Raul Palma
PSNC
rpalma@man.poznan.pl

Raphaël Troncy
EURECOM
raphael.troncy@eurecom.fr

Daniel Vila
Recogn.ai
daniel@recogn.ai

ACM Reference Format:
Carlos Badenes, Ronald Denaux, Martine De Vos, Daniel Garijo, Jose Manuel Gomez-Perez, Agnieszka Lawrynowicz, Pasquale Lisena, Raul Palma, Raphaël Troncy, and Daniel Vila. 2017. K-CAP2017 Satellites - Workshops and Tutorials. In *K-CAP 2017: K-CAP 2017: Knowledge Capture Conference CD-ROM, December 4–6, 2017, Austin, TX, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3148011.3188410

1 2ND INTERNATIONAL WORKSHOP ON CAPTURING SCIENTIFIC KNOWLEDGE (SCIKNOW 2017)

1.1 Introduction and Goals

The 2nd workshop on Capturing Scientific Knowledge (SciKnow 2017)[1] aimed to bring together researchers interested in representing and capturing knowledge about different areas of science so that it can be used by intelligent systems to support scientific research and discovery. Although great advances have been made in the last decade, scientific knowledge is still complex and poses great challenges for knowledge capture. SciKnow provided a forum to discuss existing forms of scientific knowledge representation and existing systems that use them, envisioning major areas to augment and expand this field of research.

SciKnow 2017 is the follow up event of a series which started at K-CAP 2015[2]. SciKnow 2017 took place in Austin Texas, December 4th, and had between 17 and 22 attendants during a full day event. Participants had different backgrounds, ranging from computer science and e-Science to bioinformatics and hydrology.

[1] https://sciknow.github.io/sciknow2017/
[2] https://www.isi.edu/ikcap/sciknow2015/

K-CAP 2017, December 4–6, 2017, Austin, TX, USA

ACM ISBN 978-1-4503-5553-7/17/12.
https://doi.org/10.1145/3148011.3188410

1.2 Workshop Summary

SciKnow 2017 was divided in three main sessions, which grouped seven paper presentations, together with a keynote presentation and a discussion panel. They keynote was presented by Suzanne Pierce, a hydrologist who explained the importance of decision support based methodologies to successfully communicate scientific outcomes to end users. The keynote described several examples of miscommunication, and their environmental and economic consequences. A key aspect of capturing existing scientific knowledge should be making it understandable and available to potential affected users.

The first session of the workshop focused on domain-specific knowledge capture, in particular when capturing provenance in hydrology [17] and facilitating curation of bioinformatics datasets [2]. The second session contained works related to reproducibility and reusability of scientific methods, with guidelines and metadata recommendations on how to perform abstractions on scientific experiments [18], proper design of computational notebooks [4] and spreadsheet annotation [7]. The final session dealt with the automated support for scientific processes, introducing an approach for capturing the iterative nature of scientific experiments [5] and an approach for representing hypothesis in knowledge discovery systems [9].

1.3 Discussion Summary

SciKnow 2017 was a participative discussion-oriented workshop. All presentations had several questions and comments, further elaborated at the discussion panel. Below is a summary of the main topics addressed in the workshop.

From machine in the loop systems to human in the loop systems. Nowadays human scientists are the main drivers of scientific research. Scientists formulate hypotheses, search for appropriate data,

prepare and execute experiments, collect results and write down their corresponding conclusions. Intelligent systems are used only as support tools that perform computational experiments set by scientists; or search and store results. In this context, the first topic discussed in the workshop focused on how participants envisioned the role of scientists when doing research with intelligent systems.

Participants distinguished three main roles for researchers. The first one is before an experiment takes place, guiding which sources of information and hypotheses may be most appropriate to do research on. The second one is while doing research, where scientists could interact iteratively with intelligent systems until a solution for the problem at hand is found. Finally, the last type of interaction is after an experiment is performed, in order to be able to find potential errors and curate existing results. As denoted by the roles, participants agreed for the need of humans to be kept as part of the research lifecycle. Intelligent systems may be used to help and support scientific researchers. However, researchers cannot be blind to the scientific process, otherwise they won?t be able to explain the obtained results appropriately and therefore adopt a new proposed solution.

The discussion ended by addressing the need for a better knowledge transfer and capture for intelligent systems. Existing approaches are currently able to extract information and reason with it, but they often do not understand it. Addressing this gap seems like a crucial need for intelligent systems to play more important roles in scientific research.

Real science versus ideal science. Several of the papers and examples presented at the workshop discussed case studies showcasing the complexity of real scientific problems. In computer sciences, researchers tend to simplify problems and design solutions that are generically applicable. These solutions are often difficult to adapt to real world problems in specific domains.

During the workshop we observed that current ?computer sciences approaches? mainly consider a limited fragment of the scientific process, mostly related to experiments, observations and protocols. Other, less tangible aspects of the scientific process are often left out of scope: formulation of hypotheses and research questions, assumptions and choices made by the scientists and interpretation and use of knowledge. These aspects capture the context of a research experiment, and are key for its understanding by other scientists.

In addition, supporting scientists in their work is not trivial. In order to make "computer science approaches" applicable, scientists have to be open and explicit about their data and methods. At the moment this is perceived by part of the community as risky (fear of losing ownership of data) and time consuming (due to the time required to create all metadata). Intelligent systems should highlight incentives for fostering open reusable data.

1.4 Challenges for Scientific Knowledge Capture

According to the discussions aired in the workshop, three main challenges must be tackled in order to integrate intelligent systems in the research lifecycle conducted by researchers.

The first challenge is **knowledge explainability**. Humans need to be part of the research lifecycle. Scientists should be able to intervene in situations that are surrounded by ambiguity and uncertainty, where automated reasoning may not be applicable. Besides, both scientists and non-scientific parties should be able to use the information produced by intelligent systems. In order to be understood and applied by humans, information in the research lifecycle needs to be clear, unambiguous, explicit and accompanied by semantics. Having intelligent systems help creating explanations of scientific experiments may foster their reusability.

The second challenge is **knowledge transfer**. Intelligent systems may have full access to data and source codes of a research project in order to support any research associated to it. However, access to data and software alone is not sufficient if an intelligent system cannot understand its contents or its functionality. Data, software and methods should be accompanied by metadata and information at a conceptual level. In order to address this issue, intelligent systems may also become part of the metadata annotation process, helping humans curating metadata annotations and abstractions. With these knowledge, intelligent systems would be able to help relating different datasets, software and methods together.

The third challenge is **context capture**. Scientific knowledge of research projects is only valuable when it can be understood by other scientists different from the original 'creators', and reused in different situations. This requires that the context of a research project is captured, including information on hypotheses and research questions, assumptions, design decisions, etc. that are part of the domain of interest. Context is currently captured in publications in human-readable format, but it is generally unavailable in machine readable format appropriate for intelligent systems. A way to tackle this issue may involve highlighting the benefits of such annotations, proposing assistants that use this knowledge to infer and suggest relevant lines of research.

Acknowledgements. The workshop organizers would like to thank all workshop attendees who participated in the discussions and presentations, including Yolanda Gil, Suzanne Pierce, Takahiro Kawamura, Francesco Osborne, Gully Burns, Natalia Villanueva-Rosales, Danai Symeonidou, Blake Regalia, Joe Raad, Al Idrissou, Mauro Vallati, Iacopo Vagliano, Endris Kemele, Pablo Calleja and Ruben Taelman.

2 SEMANTIC DATA MINING TUTORIAL

Semantic data mining is a data mining approach where domain ontologies are used as background knowledge [11]. The challenge is then to also mine knowledge encoded in domain ontologies and knowledge graphs in addition to purely empirical data. Data mining techniques take on input *data* and produce *models* or *pattern sets*, where the raw data may be of various forms such as a data table or text documents. The tutorial covered the topic of using *background knowledge* (also known as *prior* or *domain* knowledge) in data mining in addition to the raw data.

Semantic data mining is a relatively new research area. The term 'semantic data mining' was introduced by Kralj Novak et al. [14] in 2009. Since then, many newer approaches have been proposed [16]. The tutorial intended to provide a synthetic, unifying view on semantic data mining and its application to knowledge acquisition.

The specific aims of the tutorial were:

- to provide a synthetic, unifying view on the field,
- to present major research challenges arising from pecularities of semantic data mining,
- to demonstrate how semantic data mining can be used in practice for knowledge acquisition.

Semantic data mining becomes of more interest since knowledge acquisition becomes more and more statistical. Traditional knowledge engineering techniques are being complemented with machine learning, and machine learning methods, in turn, need to properly consume complex representations. Large amounts of data are now becoming available as structured knowledge or knowledge graphs where particular data items are described using ontologies. This availability of semantically annotated data provides both opportunities and challenges, the latter ones due to the complexity of the data, expressivity of the semantic representation languages used to represent ontologies, and uncommon assumptions made (e.g., open world). The tutorial aimed to address such challenges, including topics such as fundamentals of semantic data mining, particular data mining techniques operating directly on knowledge graphs and ontologies, and using semantic data mining for knowledge acquisition. Notably, the tutorial addressed the special topic of K-CAP'2017 dealing with statistical approaches for Web data analysis, and combining knowledge engineering and machine learning for automatic creation and refinement of knowledge graphs. The tutorial was largely based on the recently published book entitled "Semantic data mining. An ontology-based approach" [11]. It also incorporated some of the recent advances in the area, namely 'semantic' embeddings (embedding background knowledge into neural networks), and showed how to apply semantic data mining techniques to knowledge acquisition giving special attention to refinement of knowledge graphs.

The particular topics covered by the tutorial were:

- the definition of semantic data mining,
- basics of semantic data mining (data mining as search, generality relations, refinement operators),
- methods of semantic data mining (pattern mining, concept learning, similarity-based methods),
- pecularities of semantic data mining (truly 'semantic' similarity measures, dealing with the Open World Assumption),
- semantic data mining for knowledge acquisition: knowledge graph refinement (mining types, synonymous properties, disjointness, inconsistencies), mining knowledge graphs for enrichment of schemas and ontologies.

The tutorial also included a hands-on session with use of an open source tool, a plugin to Protégé, named LeoLOD Swift Linked Data Miner [15].

ACKNOWLEDGMENTS. This work has been partially supported by the Polish National Science Center (Grant No 2014/13/D/ST6/02076).

3 DOING REUSABLE MUSICAL DATA (DOREMUS)

3.1 A music data model

The DOREMUS model[3] is an extension of FRBRoo for describing cultural objects [8], applied to the specific domain of music. This is a dynamic model centered on a Work-Expression-Event triplet, which can describe different parts of the life of a work, like the Performance, the Publication or the creation of a derivative Work, each one incorporating the expression from which it comes from. DOREMUS adds specific classes and properties to FRBRoo, such as the musical key, the genre, the tempo, the medium of performance (MoP), etc. [6].

Each triplet contains an information that, at the same time, can live autonomously and be linked to the other entities. Thinking about a classic work, we will have a triplet for the composition, one for each performance event, one for every manifestation (i.e. the score), etc., all connected in the graph.

A large number of properties that are involved in the music description are supposed to contain values that are shared among different entities. These literal values can be expressed in multiple languages or in alternative forms (e.g. "sax" and "saxophone"), making reconciliation hard. Our choice is to use controlled vocabularies for those common concepts. We are using SKOS as representation model, that allows to specify for each concept the preferred and the alternative labels in multiple language, to define a hierarchy between the concepts, and to add comments and notes for describing the entity and help the annotation activity. Each concept becomes a common node in the musical graph that can connect a musical work to another, an author to a performer, etc. We collected, implemented and published 15 controlled vocabularies belonging to 6 different categories. Some of them are already available on the web, but some others were not (or not completely) published in a suitable format for the Web of Data, and some others did not exist at all so we generated them on the base of real data coming from the partners, enriched by an editorial process that involved also librarians[4].

3.2 Data Conversion

For representing music metadata, libraries make commonly use of the MARC format, which consists in a succession of fields and subfields separated by tags. Although MARC is a standard, its adoption is restricted to the library world. The benefits of moving from MARC to an RDF-based solution consist in the interoperability and the integration among libraries and with third party actors, with the possibility of realizing smart federated search [1, 3].

3.2.1 From MARC to RDF. We develop and release MARC2RDF, an open source prototype for the automatic conversion of MARC bibliographic records to RDF using the DOREMUS ontology [12]. The process relies on explicit expert-defined transfer rules that indicate where in the MARC file to look for what kind of information, as well as useful examples. The converter is composed of different modules, that works in succession. A *file* parser reads the MARC

[3] http://data.doremus.org/ontology/
[4] https://github.com/DOREMUS-ANR/knowledge-base/tree/master/vocabularies

file and makes the content accessible by field and subfield number. Then, it builds the RDF graph reading the fields and assigning their content to the DOREMUS property suggested in the transfer rules. The *free-text interpreter* extracts further information from the plain text fields, that includes editorial notes. The parsing is realized through empirically defined regular expression, that are going to be supported by Named Entity Recognition techniques as a future work. Finally, the *string2vocabulary* component performs an automatic mapping of string literals to URIs coming from controlled vocabularies. As additional feature, this component is able to recognise and correct some noise that is present in the source MARC file: as an example, this is the case of musical keys declared as genre and vice-versa.

3.2.2 Dealing With Heterogeneous Formats. Apart from MARC, we are converting other source bases (in XML), that are too specific to be handled by a single converter. Therefore, we developed *ad hoc* software that have a generic workflow: parse the input file and collect the required information, create the graph structure in RDF, run the *string2vocabulary* module described previously. This procedure creates different graphs, one for each source. Those source databases are complementary but also contain overlaps (e.g. two databases that describe the same work or the same performance with complementary metadata). We have started to automatically interlink the datasets, so that the resulting knowledge graph provides a richer description of each work.

3.2.3 Answering Complex Queries. Before the beginning of the project, a list of questions have been collected from experts of the partner institutions[5]. These questions reflect real needs of the institutions and reveal problems that they face daily in the task of selecting information from the database (e.g. concert organisation, broadcast programming) or for supporting librarian and musicologist studies. They can be related to practical use cases (the search of all the scores that suit a particular formation), to musicologist topics (the music of a certain region in a particular period), to interesting stats (the works usually performed or published together), or to curious connections between works, performances or artists. Most of the questions are very specific and complex, so that it is very hard to find their answer by simply querying the search engines currently available.

3.3 Exploration and Recommendation

We developed the first version of Overture, a web prototype of an exploratory search engine for DOREMUS data. The application makes requests directly to our SPARQL endpoint[6] and provides the information in a nice user interface. The application goal is to show a way to visualise the knowledge in DOREMUS and host a recommender system.

3.3.1 Visualizing the Complexity. At the top of the user interface, the navigation bar allows the user to navigate between the main concepts of the DOREMUS model. The challenge is in giving to the final user a complete vision on the data of each class and letting him/her understand how they are connected to each other. We keep as example Beethoven's *Sonata for piano and cello n.1* [7]. Aside from the different versions of the title, the composer and a textual description, the page provides details on the information we have about the work. When these values come from a controlled vocabulary, a link is present in order to search for expressions that share the same value. A timeline shows the most important events in the story of the work (the composition, the premiere, the first publication). Other performances and publications can be represented below. The portrait of the composer in background comes from DBpedia. It is retrieved thanks to the presence in the DOREMUS database of `owl:sameAs` links. These links comes in part from the International Standard Name Identifier (ISNI) service, in part thanks to an interlinking realised by matching the artist name, birth and death date in the different datasets.

3.3.2 Music Recommendation. What we should suggest to an user listening Beethoven? Similar musicians should share with the German composer some features: period, genre, key, casting or similar instrument played. How to define a similarity measure that involves this concepts?

We propose a solution based on graph embeddings generated at different levels: (i) For simple features (genre, key, mop), we compute for each term an embedding applying *node2vec* [10] on two sub-graphs: the one of the controlled vocabularies and the one corresponding to the usage of their values in the DOREMUS dataset; (ii) For complex features (artist), we generate the embeddings by the combination of its corresponding feature embedding; (iii) Finally, we combine simple and complex feature embeddings, following the same rules, to obtain the embedding of a work.

The use of embeddings reduces the similarity problem as the reverse of an euclidean distance. If some properties are missing, we apply a penalisation computed as percentage of missing feature in the target vector with respect to the seed one [13].

The biggest advantage of this method is that the embeddings computation is required only for the simple features, then each embedding is re-used in different combination. Moreover, different weights can be assigned to each property in order to tune up the recommendation. As future work, we plan to experiment with neural networks in order to discover the best strategy to weight the contribution of each dimension in the similarity score.

ACKNOWLEDGMENTS. This work has been partially supported by the French National Research Agency (ANR) within the DOREMUS Project, under grant number ANR-14-CE24-0020.

4 HYBRID TECHNIQUES FOR KNOWLEDGE-BASED NLP : KNOWLEDGE GRAPHS MEET MACHINE LEARNING AND ALL THEIR FRIENDS

4.1 Abstract

Many different artificial intelligence techniques can be used to explore and exploit large document corpora that are available inside organizations and on the Web. While natural language is symbolic in nature and first approaches were based on symbolic and rule-based methods (e.g., ontologies and knowledge bases), most widely

[5] https://github.com/DOREMUS-ANR/knowledge-base/tree/master/query-examples
[6] http://data.doremus.org/sparql
[7] http://overture.doremus.org/expression/614925f2-1da7-39c1-8fb7-4866b1d39fc7

used methods have been based on statistical approaches (e.g., linear methods such as support vectors machines, probabilistic topic models, and non-linear methods such as neural networks). These two approaches, knowledge-based and statistical methods, have their limitations and strengths and there is an increasing trend that seeks to combine them to get the best of both worlds. This tutorial will cover the foundations and modern practical applications of knowledge-based and statistical methods, techniques and models and their combination for exploiting large document corpora. This tutorial will first focus on the foundations of many of the techniques that can be used for this purpose, including knowledge graphs, word embeddings, neural network methods, probabilistic topic models, and will then describe how a combination of these techniques is being used in practical applications and commercial projects where the instructors are currently involved.

4.2 Audience

Researchers and practitioners both from industry and academia, as well as other participants with an interest in hybrid approaches to knowledge-based natural language processing. The tutorial is interactive, with both instructors and participants engaging in rich discussions on the topic. Some familiarity on the matter is expected but not a blocker otherwise.

4.3 Program

The tutorial comprises two main blocks, consisting of slides and hands-on exercises on jupyter notebooks. During the different blocks we touch base on different examples and applications.

4.3.1 Block 1.

- Challenges of text and natural language processing.
- Modern natural language processing methods, technologies and common tasks.
- Distributed word and feature representations, embeddings.
- Knowledge graph embeddings.
- Extending word embeddings with external knowledge: retrofitting and projection.

4.3.2 Block 2.

- Towards a vecsigrafo, bringing meaning from text into knowledge graphs.
- Evaluating vecsigrafos, visual inspection and quality assurance methods.
- Knowledge graph generation from text corpora: curation, interlinking and multilingual reuse.
- Probabilistic topic models.
- Topic-based semantic similarity

4.4 About the instructors

This tutorial is offered by the following members of the Research Lab at Expert System, Recogn.ai and Universidad Politecnica de Madrid.

Jose Manuel Gomez-Perez works in the intersection of several areas of Artificial Intelligence, including Natural Language Processing, Knowledge Discovery, Representation and Reasoning. His long-term vision is to enable machines to understand text in a way similar to how humans read, bridging the gap between both through semantically rich knowledge representations and user interfaces. At Expert System, Jose Manuel leads the Research Lab in Madrid, formed by researchers, software engineers and linguists, in the belief that such vision is best served through a combination of structured knowledge graphs and probabilistic approaches. Before Expert System, he worked at Intelligent Software Components, one of the first European companies to deliver Semantic and Natural Language Processing solutions on the Web. He also consults for companies like Coca-Cola, British Telecom, Volkswagen, HAVAS and ING. Also active as an entrepreneur, he co-founded a startup and advised another. An ACM member and former Marie Curie fellow, Jose Manuel holds a Ph.D. in Computer Science and Artificial Intelligence and regularly publishes in top scientific conferences and journals in the area. His views on AI and its applications have appeared in magazines like Nature and Scientific American. In 2015, Jose Manuel was the program chair of K-CAP, the International Conference on Knowledge Capture.

Ronald Denaux is a senior researcher at Expert System Iberia. Ronald obtained his MSc in Computer Science from the Technical University Eindhoven, The Netherlands. After a couple of years working in industry as a software developer for a large IT company in The Netherlands, Ronald decided to go back to academia. Ronald obtained a PhD, again in Computer Science, from the University of Leeds, UK. Ronald's research interests have revolved around making semantic web technologies more usable for end users, which has required research into (and resulted in various research publications in) the areas of Ontology Authoring and Reasoning, Natural Language Interfaces, Dialogue Systems, Intelligent User Interfaces and User Modelling. Besides research, Ronald has recently also been involved in knowledge transfer and product development from research prototypes.

Daniel Vila is co-founder of recogn.ai, a Madrid-based startup and spin-off from Universidad Politecnica de Madrid, building next generation solutions for text analytics and content management using the AI methods. Daniel holds a PhD in Artificial Intelligence by Universidad Politecnica de Madrid (2016), where he worked at the Ontology Engineering Group and developed the solution supporting a large knowledge graph combining NLP and semantic technologies: the datos.bne.es data service from the National Library of Spain.

Carlos Badenes: After more than 8 years working on the M2M world, Carlos began researching about text mining within the context of semantic web. Since then, he has moved more deeply into the study of topic modeling techniques to analyze large collections of documents, incorporating semantic resources and working on multilingual domains. He currently works as an associate researcher at the Ontology Engineering Group (OEG) doing a PhD at the Universidad Politecnica de Madrid (UPM).

4.5 Materials

If interested in the materials of the tutorial, please contact Jose Manuel Gomez-Perez at `jmgomez@expertsystem.com`

ACKNOWLEDGMENTS. Partially funded by the EU H2020 and national research projects xLiMe-ES (20160805) and DANTE (700367).

REFERENCES

[1] Getaneh Alemu, Brett Stevens, Penny Ross, and Jane Chandler. 2012. Linked Data for libraries: Benefits of a conceptual shift from library-specific record structures to RDF-based data models. *New Library World* 113, 11/12 (2012), 549–570.

[2] Gully A Burns, Randi Vita, James Overton, Ward Fleri, and Bjoern Peters. 2017. Semantic Modeling for Accelerated Immune Epitope Database (IEDB). In *Proceedings of Workshops and Tutorials of the 9th International Conference on Knowledge Capture (K-CAP2017)*.

[3] Gillian Byrne and Lisa Goddard. 2010. The strongest link: Libraries and linked data. *D-Lib magazine* 16, 11 (2010), 5.

[4] Lucas A. M . C. Carvalho, Regina Wang, Yolanda Gil, and Daniel Garijo. 2017. NiW: Converting Notebooks into Workflows to Capture Dataflow and Provenance. In *Proceedings of Workshops and Tutorials of the 9th International Conference on Knowledge Capture (K-CAP2017)*.

[5] Lucas A. M. C. Carvalho, Daniel Garijo, Bakinam T. Essawy, Claudia Bauzer Medeiros, and Yolanda Gil. 2017. Requirements for Facilitating the Continuous Creation of Scientific Workflow Variants. In *Proceedings of Workshops and Tutorials of the 9th International Conference on Knowledge Capture (K-CAP2017)*.

[6] Pierre Choffé and Françoise Leresche. 2016. DOREMUS: Connecting Sources, Enriching Catalogues and User Experience. In *24^{th} IFLA World Library and Information Congress*. Colombus, USA.

[7] Martine de Vos, Jan Wielemaker, Bob Wielinga, Guus Schreiber, and Jan Top. 2017. How plausible is automatic annotation of scientific spreadsheets?. In *Proceedings of Workshops and Tutorials of the 9th International Conference on Knowledge Capture (K-CAP2017)*.

[8] Martin Doerr, Chryssoula Bekiari, and Patrick LeBoeuf. 2008. FRBRoo: a conceptual model for performing arts. In *CIDOC Annual Conference*. Athens, Greece, 6–18.

[9] Daniel Garijo, Yolanda Gil, and Varun Ratnakar. 2017. Capturing Hypothesis Evolution in Automated Discovery Systems. In *Proceedings of Workshops and Tutorials of the 9th International Conference on Knowledge Capture (K-CAP2017)*.

[10] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks. In *22^{nd} ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, USA.

[11] Agnieszka Ławrynowicz. 2017. *Semantic data mining. An ontology-based approach*. Studies on the Semantic Web, Vol. 29. IOS Press.

[12] Pasquale Lisena, Manel Achichi, Eva Fernandez, Konstantin Todorov, and Raphaël Troncy. 2016. Exploring Linked Classical Music Catalogs with OVERTURE. In *15^{th} International Semantic Web Conference (ISWC)*. Kobe, Japan.

[13] Pasquale Lisena and Raphaël Troncy. 2017. Combining Music Specific Embeddings for Computing Artist Similarity. In *18^{th} International Conference on Music Information Retrieval (ISMIR), Late-Breaking Demo Track*. Suzhou, China.

[14] Petra Kralj Novak, Anze Vavpetic, Igor Trajkovski, and Nada Lavrac. 2009. Towards semantic data mining with g-SEGS. In *Proceedings of the 11th International Multiconference Information Society (IS 2009)*, Vol. 20.

[15] Jedrzej Potoniec and Agnieszka Lawrynowicz. 2016. A Protege Plugin with Swift Linked Data Miner. In *Proceedings of the ISWC 2016 Posters & Demonstrations Track co-located with 15th International Semantic Web Conference (ISWC 2016), Kobe, Japan, October 19, 2016*. http://ceur-ws.org/Vol-1690/paper48.pdf

[16] Petar Ristoski and Heiko Paulheim. 2016. Semantic Web in data mining and knowledge discovery: A comprehensive survey. *J. Web Sem.* 36 (2016), 1–22. https://doi.org/10.1016/j.websem.2016.01.001

[17] Natalia Villanueva-Rosales, Luis Garnica Chavira, Smriti Rajkarnikar Tamrakar, Deana Pennington, Raul Alejandro Vargas-Acosta, Frank Ward, and Alex S. Mayer. 2017. Capturing Scientific Knowledge for Water Resources Sustainability in the Rio Grande Area. In *Proceedings of Workshops and Tutorials of the 9th International Conference on Knowledge Capture (K-CAP2017)*.

[18] Daniel Garijo Yolanda Gil, Margaret Knoblock, Alyssa Deng, Ravali Adusumilli, Varun Ratnakar, and Parag Mallick. 2017. A Workflow Design Methodology to Improve Reproducibility andReusability of Computational Experiments. In *Proceedings of Workshops and Tutorials of the 9th International Conference on Knowledge Capture (K-CAP2017)*.