Informacijska družba IS 2007 / Information Society IS 2007
10. mednarodna multi-konferenca / 10th International Multi-Conference
Vzgoja in izobraževanje v informacijski družbi / Education in Information Society
Ljubljana, Slovenija, 12. oktober 2007

Structuring Domain Knowledge by Semi-automatic Ontology Construction

Bojan Cestnik [1,3], Ingrid Petrič [2], Tanja Urbančič [2], Marta Macedoni-Lukšič [4]

[1] Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

[2] University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia

[3] Temida, d.o.o., Dunajska cesta 51, 1000 Ljubljana, Slovenia

[4] University Children's Hospital, University Medical Center, Ljubljana, Slovenia

bojan.cestnik@temida.si, ingrid.petric@p-ng.si, tanja.urbancic@p-ng.si, marta.macedoni-luksic@mf.uni-lj.si

Abstract

In this paper we present a case in semi-automatic ontology construction from literature. In this case we concentrate on the articles about autism obtained from PubMed Central database. Our motivation was to investigate how separate parts of articles, such as titles, abstracts and full texts, influence the constructed ontology. Our results confirm the intuitive expectation that constructing ontologies from abstracts is a rational choice when uncovering the structure of a given scientific field. Also, when compared to the general autism knowledge, ontology concepts from abstracts show the highest resemblance.

Keywords: *Knowledge management, education, concept learning, ontologies, autism*

1. Introduction

For the whole era of science, ontologies have been used as a means to organize scientific information and, more importantly, to provide a common vocabulary of concepts for the education process. Until recently, the practice of ontology construction relied mostly on manual extraction of interesting concepts from scientific literature and organizing them in a suitable hierarchy. Nowadays, the largely increased amount of scientific publications requires automated support for the task. With new knowledge technologies, selected scientific articles can be processed semi-automatically and, therefore, the process of ontology construction made more effective and feasible in practice.

Ontologies play a substantial role in the education process (e.g. Breuker et al., 1999). Although content has always been considered a crucial factor in education, the emphasis in educational research has been also on form. In this perspective, ontologies in education are parts of the common sense understanding of the world that define the concepts and structures in a domain. So, ontologies are particularly important when the education process embraces not only skill acquisition but also insight and understanding.

In information science, ontology is a data model that represents a domain and is used to reason about the objects in that domain and the relations between them. So, ontologies have the capability to share a common understanding of domains, and therefore to support a research with ability to reason over and to analyze the information at issue (Joshi and Undercoffer, 2004). In recent years, many tools that help constructing ontologies from texts in a given problem domain were developed and successfully used in practice (Brank et al., 2005). Among them, OntoGen (Fortuna et al., 2006) received a notable attention in the text-mining community.

Nowadays, researchers and students are faced with vast amounts of data when extracting knowledge from the rapidly growing volumes of databases. The situation becomes even more striking when a person wants to obtain an insight in a field that does not fall directly into her or his area of expertise. There is a special field of knowledge discovery in databases that aims at supporting researchers and students for such tasks. Knowledge discovery is the process of discovering useful knowledge from data, which includes data mining as the application of specific algorithms for extracting patterns from data (Fayyad et al. 1996). In fact, important information hidden in huge databases could be revealed by data mining and knowledge discovery techniques. When databases contain bibliographic semi-structured data, text mining as a specific type of data mining can be used.

When constructing ontologies from scientific articles in a semi-automatic manner there is a decision to be made: which parts of articles to include in the process. While some experts suggest that the more text you can obtain, the better the constructed ontology, others advocate a more systematic approach that relies on comparably balanced parts of explored texts (Cohen et al., 2005). With the experiments described in this article we wanted to clarify this dilemma. So, our main motivation was to analyze how separate parts of scientific articles influence the constructed ontologies. Initial results presented in (Petrič et al., 2006) encouraged further investigation that enabled us to present our findings in a more systematic fashion. When evaluating which parts of articles are more appropriate for the ontology construction, we assessed two criteria: first, the pair-wise similarity of the constructed ontology concepts, and second, their resemblance to the commonly accepted concepts in a given domain.

The set of articles for our study was selected from the autism domain. Autism belongs to a group of pervasive developmental disorders that are portrayed by an early delay and abnormal development of cognitive, communication and social interaction skills of a person. Heterogeneity of this developmental disturbance and its different degrees of affecting children has led to contemporary naming of autism with term Autism spectrum disorders (ASD). Due to its rather complex nature the domain still lacks a thorough understanding of the underlying phenomena and, therefore, further investigations are needed (Persico and Bourgeron, 2006). Our team is active in investigations towards finding new methods for early diagnosis in autism. We are particularly focused on extracting knowledge from vast amounts of textual data and presenting it in a human readable form that helps to get better insight and understanding of the domain (Urbančič at al., 2007).

This paper is organized as follows. First, we give a short overview of ontology construction approaches. In section 3 we present our experiments on documents about autism. Section 4 contains the evaluation of the obtained ontologies. The most important findings are summarized in the conclusion.

2. Semi-automatic ontology construction

Ontologies are used in information science as a form of knowledge representation about the world or some part of it. In general, ontologies include descriptions of objects, concepts, attributes and relations between objects. Traditionally, ontologies for a given domain are constructed manually using some sort of language or representation and rely on manual extraction of common sense knowledge from various sources. Recently, several programs that support manual ontology construction have been developed, like for example Protégé (Gennari, 2002).

Since manual ontology construction is a complex and demanding process, there is a strong tendency to provide a computerized support for the task. Based on text mining techniques that have already proven successful for the task, OntoGen (Fortuna et al., 2006) is a tool that enables interactive construction of ontologies from text documents in a selected domain. A user can create concepts, organize them into topics and also assign documents to concepts. With the use of machine learning techniques OntoGen supports individual phases of ontology construction by suggesting concepts and their names, by defining relations between them, and by automatic assignment of documents to the concepts (Fortuna, 2006).

Our main motivation for using OntoGen was to gain a quick insight into a given domain by semi-automatically generating the main ontology concepts from the domain's documents. The semi-automatic ontology construction method implemented in OntoGen incorporates basic text mining principles. The input for the tool is a collection of text documents. Documents are represented as vectors, which is often referred as vector-space model. Using this representation, similarity between two documents can be defined as the cosine of the angles between the two corresponding vector representations. When suggesting new concepts, OntoGen uses *K*-means clustering technique (Jain et al., 1999) and keyword extraction method (Brank et al., 2002).

3. Experiments on documents about autism

For the purpose of this analysis we decided to use professional literature that is publicly accessible on the Internet in the PubMed database of biomedical publications. In this database we found 10.821 documents (till August 21, 2006) that contain words in the form of autis*, which we used as the search criterion for the articles about autism. Documents were prevailingly described with titles, authors and abstracts. However, there were 354 articles that were presented in the database also with the entire text. Other relevant publications were either restricted to abstracts of documents or their entire texts were published in sources outside PubMed. From the listed 354 articles we further restricted the target set of articles to those that have been published in the last ten years. As a result, we ended up with 214 articles from 1997 forward. To use them in our experiments, we partitioned them to titles, abstracts and texts.

3.1 Design of experiments

When designing the experiments, we had two goals in mind. First, we wanted to get acquainted with the domain in the sense that we better understand the underlying concepts. Second, we wanted to evaluate different ontologies constructed on different parts of documents, such as titles, abstracts and texts. In addition, we tried to evaluate also the content compliance between titles, abstracts and entire bodies of texts of the related documents. Besides, we wanted also to experiment with different values of parameter *k* that is used by OntoGen's *K*-means clustering algorithm.

From the 214 documents obtained by our search in the PubMed Central data base we created three input text files: the file with 214 titles, the file with 214 abstracts, and the file with 214 bodies of texts without their respective titles and abstracts. Each text file was used separately as an input for OntoGen; in the process of semi-automatic ontology construction we used OntoGen to construct several top-level ontology concepts and describe them with suggested keywords. The ontologies were built with two values for parameter *k*: first, with parameter *k*=8 that was automatically suggested by OntoGen, and second, with parameter *k*=5 that experimentally turned out to be a well-balanced tradeoff between the complexity and comprehensibility in this domain. Moreover, the results obtained with *k*=5 were more in accordance with the concepts found in the autism survey literature (Zerhouni, 2004) and were also evaluated well by an expert from autism domain.

In this way OntoGen generated 8 and 5 concepts respectively on the first level of domain ontology for each of the input files (titles, abstracts and bodies of texts). Each concept was described with three most relevant keywords as suggested by Ontogen. Our evaluation of the obtained ontology concepts was first performed on vocabulary level comparing keywords of different concepts and analyzing the sets of documents that correspond to each concept. Next, concept descriptions were presented to the medical expert that also evaluated the concepts from her perspective.

3.2 Experimental results

In this subsection we present the results of our experiments. Each table from Table 1 to Table 6 contains ontology concepts described with three keywords (labeled Keywords) and the number of related documents (labeled No. Docs).

Id	Keywords	No. Docs
0	Root	214
1	preference, assessment, effects	31
2	reinforcement, children_autism, early	27
3	genes, susceptibility, specific	32
4	functioning, syndrome, analysis	26
5	autism, teach, child	25
6	vaccination, schedules, activated	24
7	social, evidence, chromosome	17
8	disorders, linkage, case	32

***Table 1**: 8 concepts of autism ontology generated from 214 titles.*

Id	Keywords	No. Docs
0	Root	214
1	sensory, sounds, auditory	8
2	stereotypy, behavioral, probl_beha	26
3	reinforcers, preferred, stimulus	41
4	teach, question, procedure	18
5	gene, linkage, regional	60
6	parent, mmr, vaccine	16
7	language, age, children	28
8	vaccine, mmr, mmr_vaccine	17

***Table 2**: 8 concepts of autism ontology generated from 214 abstracts.*

Id	Keywords	No. Docs
0	root	214
1	executive, nv, cortical	26
2	stereotypies, reinforcement, prob_be	27
3	reinforcement, session, aggression	38
4	prompted, script, teaching	21
5	linkage, family, gene	55
6	ht, secretin, legs	8
7	chemical, infant, sleep	14
8	vaccine, mmr, mmr_vaccine	25

***Table 3**: 8 concepts of autism ontology generated from 214 bodies of texts.*

Id	Keywords	No. Docs
0	root	214
1	autism, children_autism, children	67
2	syndrome, detection, social	19
3	disorders, spectrum, neurodevelopmental	39
4	genetic, chromosome, linkage	50
5	reinforcement, effects, behavior	39

***Table 4**: 5 concepts of autism ontology generated from 214 titles.*

Id	Keywords	No. Docs
0	Root	214
1	reinforcers, behavioral, problems_behavioral	49
2	language, foxp2, children	52
3	reinforcers, vaccine, aggression	46
4	linkage, gene, regional	55
5	virus, infection, trim5alpha	12

Table 5*: 5 concepts of autism ontology generated from 214 abstracts.*

Id	Keywords	No. Docs
0	root	214
1	reinforcement, session, trial	72
2	reinforcement, sleep, infant	37
3	vaccine, mmr, mmr_vaccine	24
4	linkage, family, gene	71
5	infection, pml, patients	10

Table 6*: 5 concepts of autism ontology generated from 214 bodies of texts.*

4. Evaluation of the obtained ontologies

In most cases ontologies are rather complex structures. Therefore, it is often more reasonable to focus the attention on the evaluation of separate levels of ontology, rather than on the direct evaluation of whole ontologies (Brank et al., 2005). In our comparison of the ontology concepts from autism, built with OntoGen, we focused on vocabulary level of the obtained concept descriptions and related concept documents. We observed the distribution of documents within individual ontology groups on the first level of each ontology model (first level subgroups of autism domain), considering terminology that was selected by OntoGen for the presentation of concepts.

4.1 Ontology concepts from various parts of texts

The distribution of documents among 8 concepts of titles ontology (Table 1) is quite uniform. Contrary to that, the ontologies of 8 abstract concepts (Table 2), and 8 text concepts (Table 3) both show one major subgroup of documents that treat genetics, and another important group that describes reinforcers or stimulus for autists. Documents distributions in ontologies of 5 subgroups are a little different. There are two major groups of titles (Table 4) and texts (Table 6). The biggest group of titles describes autism in general, whereas the largest text group writes about reinforcement trials. The second major group in both cases (titles and texts) deals with genetics. Abstracts distributions (Table 5), on the contrary, show two most important groups that both treat different aspects of genetics.

While the first group is described with clear genetic keywords, the second group includes, among others, the keyword *foxp2*, which is a gene important for the development of speech.

The evaluation of the obtained results showed considerable differences between ontology concepts constructed from titles, abstracts and related bodies of texts. Figure 1 shows the result of comparison of 5 ontology concepts generated from abstracts and entire bodies of texts. The major similarity is identified between the groups of genetic documents, which include the same 51 articles from the observed dataset. Also, relatively large similarity can be seen also between the text and abstract groups that deal with virus infections. Also, there is a similarity between the group of texts and corresponding abstracts' subgroup about vaccine. Slightly less specific is similarity between abstracts and texts from groups: reinforcement, session, trial and reinforcement, sleep, infant. Although the concept matching presented in Figure 1 is not completely evident, a general tendency can be well spotted in the diagonal elements.

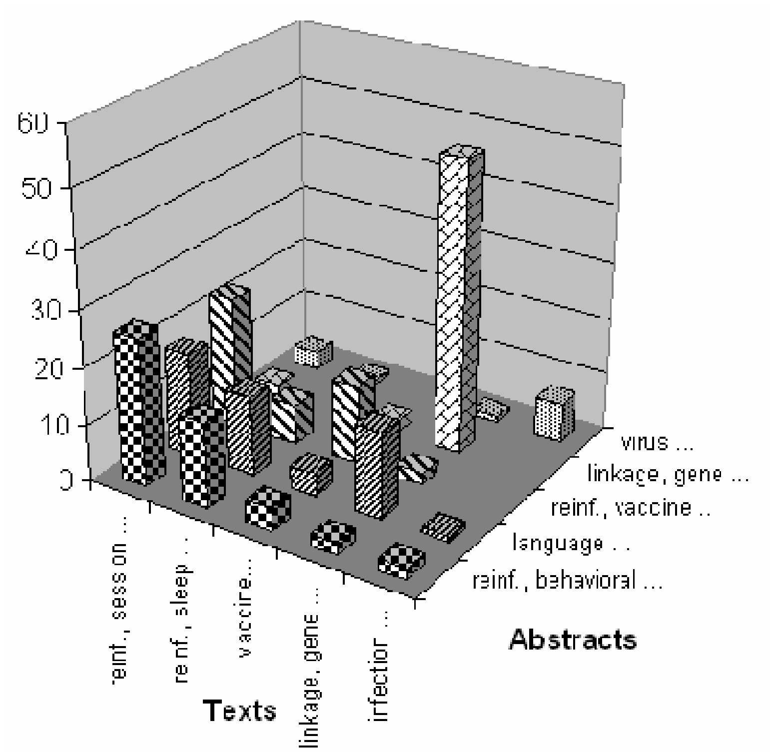

Figure 1*: Comparison between the distributions of documents belonging to the ontology concepts of abstracts and bodies of texts.*

Compared to analysis of matching bodies of texts and abstracts, we observed significantly lower similarity between texts and titles of the related articles, as well as between their abstracts and titles. Articles about genetics are the only fairly important group of documents that apparently use more similar vocabulary in titles, abstracts and entire bodies of texts. The likely cause for this observation lies in the genetic terminology and in the genetic context itself, which is quite specific when compared to other fields of autism research.

The obtained high-level concepts were presented to the expert from autism domain. She found the tables informative and in accordance with her line of reasoning in autism. In particular, the clustering of the selected articles was in most cases quite intuitive, although the keyword description of some of the generated concepts was not so straightforward. Important confirmation of the resulted ontology construction is also the recent state of autism research as described by Zerhouni (2004) that summarizes the main scientific activities of autism research in the major areas of epidemiology, genetics, neurobiology, environmental factors and specific treatments of autism. As advocated by

OntoGen's literature (Fortuna, 2006), we renamed the concepts accordingly, based on the suggested keywords. The resulted ontology concepts are presented in Figure 2.

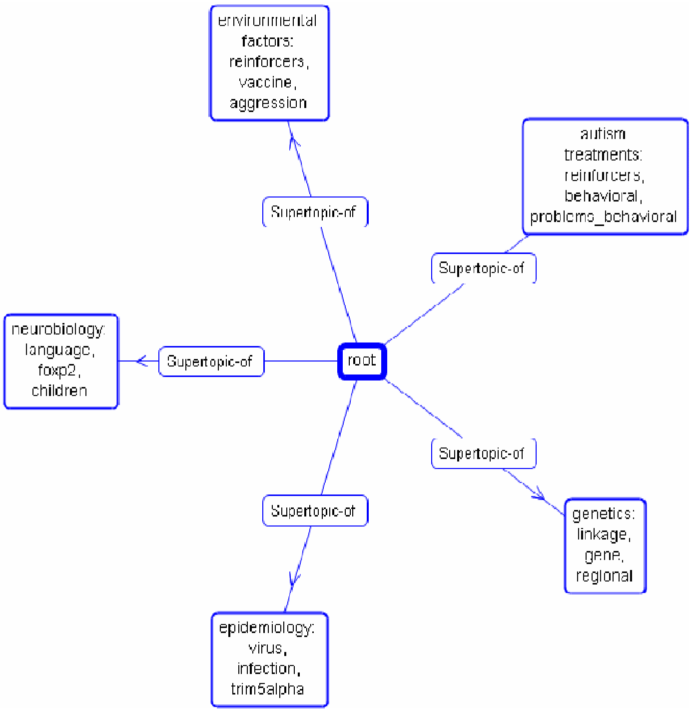

Figure 2*: Top-level autism ontology concepts as suggested by Ontogen and renamed according to autism survey literature. Original concept descriptions are included for easier identification.*

4.1 Ontology concepts with various values of k

Clustering algorithms, such as *K*-means clustering, are useful tools for data mining; however when we have to cluster datasets, it is not always clear which is the most appropriate number of clusters (parameter *k*) to use (Jain et al., 1999). OntoGen automatically proposes to use 8 clusters as a default. However, it is strongly recommended to experiment also with different *k* in order to find out the best result for the investigated domain.

After experimenting with OntoGen's default parameter *k*=8, we also constructed top-level ontology concepts with several other values for parameter *k* ranging from 2 to 15. As a result we found out that the value 5 for *k* represents a well-balanced tradeoff between the complexity and comprehensibility of the single level ontology concepts in this domain. Although the concepts generated with other values of parameter *k* also revealed some interesting domain properties, they were either too broad when *k* was small or too narrow when *k* was too large. Therefore, a careful selection of parameter *k* is very important prerequisite when constructing ontologies in a semi-automatic way.

5. Conclusion

Using tools for semi-automatic ontology construction from scientific articles can significantly speed up the process of getting acquainted with the domain of interest. Instead of reading piles of literature researchers and students can first generate top-level domain ontology concepts and thus obtain a bird's view and understanding of the domain. After that, a detailed study of the concepts of interest might be in order. In such way, semi-automatically constructed ontologies actually helped us to review and understand the complex and heterogeneous spectrum of scientific articles about autism.

Our next motivation was to investigate how separate parts of articles, such as titles, abstracts and full texts, influence the constructed ontology. In this comparison we decided to take into account only the top-level ontology concepts, mostly because comparing full-scale ontologies can become a very intricate task (Brank et al., 2005). Our graphic presentation of compared ontologies clearly exposes the main clusters of autism articles, which are shown as the highest columns in the graph in Figure 1. This way it provides a powerful way to visualize the most important similarities between observed ontologies; it can be seen that the largest collection of autism documents always deal with genetics.

Determining the proper number of top-level concepts (value of parameter *k*) for a specific domain is very important when constructing ontologies in a semi-automatic way. The goal is to find a well-balanced tradeoff between the complexity and comprehensibility of the single level ontology concepts in the domain. However, experimenting with other values of parameter *k* may also reveal some interesting domain properties.

The experimental result show that there is a substantial similarity between constructed ontology concepts from abstracts and full texts, while there is less similarity between ontology concepts from titles and abstracts and titles and full texts. This finding suggests that titles are not informative enough to be taken as the only source for constructing ontologies.

Compared to the general autism knowledge, ontology concepts from abstracts show the highest resemblance. Our results confirm the intuitive expectation that constructing ontologies from abstracts is a rational choice when uncovering the structure of a given scientific field. The titles as well as full texts are typically less useful for the given task. When dealing with full texts, some preprocessing tasks like stemming and stopping can improve the utility (Cohen et al., 2005).

For further work we consider experimenting also with sources that are mixtures of titles, abstracts and full texts. Often there are some articles available as abstract-only and some other articles as full texts. A practical question that we would like to investigate is the following: is it wise to join the two sets unaltered or is it better to include also the latter set in abstract-only form.

6. Acknowledgement

This work was partially supported by the Slovenian Research Agency programme Knowledge Technologies (2004-2008). We wish to thank Nada Lavrač for her suggestion to use OntoGen and Blaž Fortuna for his discussions about OntoGen's performance.

References

Breuker, .J., Muntjewerff, A., and Bredeweg, B., (1999): »Ontological modeling for designing educational systems«. Proceedings of the Workshop on Ontologies for Intelligent Educational Systems, AI-ED 99, Le Mans, France, July 18-19.

Brank, J., Grobelnik, M., Milic-Frayling, N., and Mladenić, D., (2002): »Feature selection using support vector machines«. Proceeding of the Third International Conference on Data Mining Methods and Databases for Engineering, Finance, and Other Fields, Bologna, Italy, September 25-27, 2002.

Brank, J., Grobelnik, M., and Mladenić, D., (2005): »A survey of ontology evaluation techniques«. In: SIKDD 2005 at multiconference IS 2005, Ljubljana, Slovenia, October 17, 2005.

Cohen, A. M., Yang, J., Hersh, W.R., (2005): »A Comparison of Techniques for Classification and Ad Hoc Retrieval of Biomedical Documents«. In: Proceedings of the Fourteenth Annual Text REtrieval Conference, TREC 2005, Gaithersburg, MD

Fortuna, B., (2006): [http://ontogen.ijs.si/index.html], OntoGen: Description.

Fortuna, B., Grobelnik, M., and Mladenić, D., (2006): »System for semi-automatic ontology construction«. Demo at ESWC 2006. Budva, Montenegro, June, 2006.

Gennari, J., Musen, M. A., Fergerson, R. W., Grosso, W. E., Crubezy, M., Eriksson, H., Noy, N. F., and Tu, S. W., (2002): »The Evolution of Protégé: An Environment for Knowledge-Based Systems Development«.

Jain, A. K., Murty, M. N., and Flynn, P. J., (1999): »Data Clustering: A Review«. ACM Computing Surveys, Vol. 31/3, pp. 264-323.

Joshi, A. and Undercoffer, J.L., (2004): »On Data Mining, Semantics, and Intrusion Detection. What to Dig for and Where to Find It«. In: Data mining. Next Generation Challenges and Future Directions. Menlo Park, California. pp. 437-460.

Persico, A. M., Bourgeron, T., (2006): »Searching for ways out of autism maze: genetic, epigenetic and environmental clues«. Trends in neurosciences, Elsevier, Vol. 29, No. 7, July 2006.

Petrič, I., Urbančič, T. and Cestnik, B., (2006): Comparison on ontologies built on titles, abstracts and entire texts of articles. Proceedings of the 9th International multi-conference Information society IS-2006, Ljubljana, Slovenia, pp. 227-230.

Urbančič, T., Petrič, I., Cestnik, B., and Macedoni-Lukšič, M., (2007): »Literature Mining: Towards Better Understanding of Autism«. In Artificial Intelligence in Medicine LNAI 4594 (R. Bellazzi, A. Abu-Hanna, J. Hunter, eds.), Springer, 2007, pp. 217-226.

Zerhouni, E. A., (2004): »Congressional Appropriations Committee Report on the State of Autism Research«. Report for National Institutes of Health and National Institute of Mental Health, Department of Health and Human Service, Bethesda, Maryland, 2004.

About authors

Bojan Cestnik *is the general manager of software company Temida and a researcher at the Department of Knowledge Technologies at the Jožef Stefan Institute in Ljubljana. He obtained his PhD in Computer Science at the Faculty of Electrical Engineering and Computer Science, University of Ljubljana. His professional and research interest include knowledge based information systems and machine learning. His research work was presented at several international conferences. He has been involved in several large-scale software development and maintenance projects.*

Ingrid Petrič *is teaching assistant for the courses of Computer science and Business information systems at the University of Nova Gorica. She is also a doctoral student of New media and e-science at the Jožef Stefan International Postgraduate School in Ljubljana. She received her MSc in Information Management at Faculty of Economics, Univeristy of Ljubljana in 2004. As a research fellow at the Center of Systems and information technologies at the University of Nova Gorica, she investigates the available data about autism spectrum disorders by using data mining and decision support tools.*

Tanja Urbančič *is Dean of the School of Engineering and Management at the University of Nova Gorica. She is also research fellow at the Department of Knowledge Technologies at the Jožef Stefan Institute. Se received her PhD in Computer Science from the Univeristy of Ljubljana. Her current professional and research interest is mainly in knowledge management, especially in its applications to education, medicine and public health. She is a co-author of several book chapters and journal papers in IEEE Transactions on SCM, Engineering Applications of Artificial Intelligence, Journal of Intelligent and Fuzzy Systems and others.*