Towards End-to-End Transformation of Arbitrary Tables from Untagged Portable Documents (PDF) to Linked Data

Alexey Shigarov[1,2], Igor Cherepanov[1,2], Evgeniy Cherkashin[1,2],
Nikita Dorodnykh[1], Vasiliy Khristyuk[1], Andrey Mikhailov[1],
Viacheslav Paramonov[1,2], Egor Rozhkow[1,2], and Alexandr Yurin[1]

[1] Matrosov Institute for System Dynamics and Control Theory, Siberian Branch of Russian Academy of Sciences, Lermontov St. 134, Irkutsk, Russia
shigarov@icc.ru
[2] Institute of Mathematics, Economics and Informatics, Irkutsk State University, Gagarin Blvd. 20, Irkutsk, Russia

Abstract. The paper is devoted to the problem of an end-to-end table transformation from untagged portable documents (PDF) to linked data. It covers the issues of the table extraction from documents, the reconstruction of logical table structure, the conceptualization of their natural-language content, and the linking of extracted data with external vocabularies. We consider some perspective approaches for the deep-learning-based table detection, heuristic-based table structure recognition, rule-based table analysis, and knowledge-based table interpretation. They can be used as a basis to develop a consistent solution for this problem. Our application experience confirms that such solutions are demanded for populating databases and generating ontologies with tabular data being extracted from weakly and semi-structured documents.

Keywords: Data transformation · Document analysis · Table extraction · Table analysis · Table interpretation · Linked data.

1 Introduction

PDF[3] (Portable Document Format) is a very popular way to represent non-editable documents. Many of PDF documents are machine-readable but remain untagged (i.e. there are no tags for identifying layout items such as paragraphs, columns, or tables). Such documents often contain tables with arbitrary layout (e.g. cross-tabulations, invoices, and data sets). These tables can be a valuable source in various applications of the text and data analysis. However, difficulties that inevitably arise with the extraction and integration of the tabular data presented in untagged PDF documents often hinder the intensive use of them in practice.

[3] https://www.iso.org/standard/63534.html

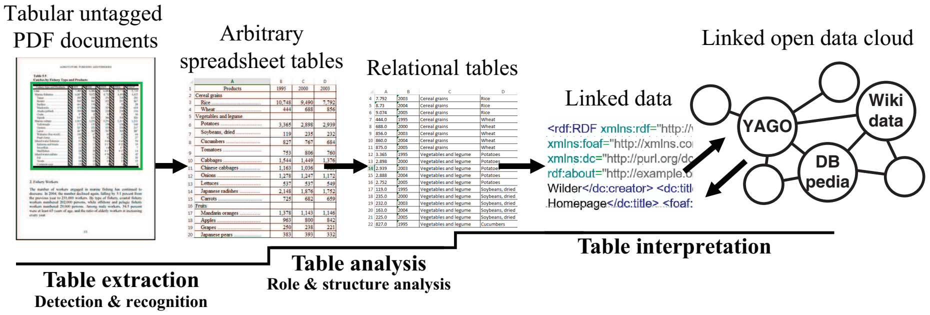

Tabular untagged PDF documents
Arbitrary spreadsheet tables
Relational tables
Linked data
Linked open data cloud
<rdf:RDF xmlns:rdf="http://
xmlns:foaf="http://xmlns.co
xmlns:dc="http://purl.org/dc
rdf:about="http://example.
Wilder</dc:creator> <dc:titl
Homepage</dc:title> <foaf:
YAGO
Wiki data
DB pedia
Table extraction
Detection & recognition
Table analysis
Role & structure analysis
Table interpretation

Fig. 1. The tasks for converting tabular data from untagged PDF documents to linked data.

The untagged PDF documents do not provide any metadata for describing table location and cell positions, as well as functional roles, internal and external relationships of data items. In other words, there is no explicit semantics that is needed for the interpretation of tabular data presented in such documents by third-party software applications. The generic approach to make the tabular data interpretable is to extract them from the documents and represent them as *linked data*. The transformation of the tabular data requires recovering the metadata missed in original documents and linking the extracted data with external vocabularies. This process can be considered as a chain of tasks as shown in Fig. 1 where each step enriches the tabular data with explicit semantics.

We present some perspective approaches to the development of a consistent solution to the end-to-end table transformation from weakly and semi-structured documents to linked data. They cover deep-learning-based table detection (identifying positions and content of tables), heuristic-based table structure recognition (identifying positions and content of cells), rule-based table analysis (recovering functional data items and their relationships), and knowledge-based table interpretation (linking recovered data items with external vocabularies). Our application experience confirms that such solutions are demanded for populating databases and generating ontologies with tabular data being extracted from weakly and semi-structured documents.

2 Table Extraction

The first task (*table extraction*) aims at extracting tables from a document and representing them in a spreadsheet-like format. This task consists of two consecutive sub-tasks: *table detection* and *table structure recognition*. The *table detection* extracts positions of a table in a document, i.e. either geometric positions of a bounding box of the table or a set of all text chunks placed inside this bounding box. The *table structure recognition* recovers positions and content of cells in a table. A detected table should be separated into cells addressed by rows and columns. The recovered cells of the extracted table can be represented as arbitrary tables in a spreadsheet format (e.g. Excel).

Many table extraction methods are traditionally heuristics-based [62, 40, 43, 42]. Some of them combine heuristics and machine-learning approaches [4, 22, 41]. The current trend involves deep learning techniques: binary classification based on convolutional neural networks [29], fine-tuned object detection models [47, 26, 2, 58], and semantic segmentation [30, 33]. Some of the existing methods show a high accuracy on the competition datasets (UNLV[48], Marmot[4], ICDAR 2013 [27], and ICDAR 2017 [25]). However, the complexity of a document layout often prevents the applicability of these methods in practice. For example, a long table placed on several consecutive pages cannot be extracted by these techniques as a whole object. We believe that the research of issues of the table extraction is far from being completed.

We propose to combine deep learning and heuristic-based techniques to develop a comprehensive solution for this task. As contemporary works [47, 26, 2, 58] show the deep neural network models provide a good performance for the table detection. The generic approach to develop such models is based on fine-tuning pre-trained models for object detection on images. We adopted this approach to create and use own model for the table detection.

To examine this approach we utilized the well-known architecture "Faster R-CNN" [44] with the pre-trained convolutional neural network model (ResNet-101) for the feature extraction. It also used the distance transformation described in [26] for pre-processing images. The regional-proposal network model was trained on three existing datasets of documents (UNLV, Marmot, and ICDAR 2017) containing 2800 samples. Additionally, the training data were augmented by affine transformations that allowed us to improve the accuracy of table detection by 5%. The performance evaluation of the developed model showed a high recall (0.979) and precision (0.865) on the ICDAR 2013 competition dataset [27]. It should be noted that the verification of the obtained predictions can decrease false positives and improves precision.

At the stage of the cell structure recognition, we propose to use our heuristic-based bottom-up method [55, 54]. Its main idea consists in building text blocks from words placed inside the bounding box of a detected table. Each text block is a whole textual content of one cell in this table. Our method exploits a set of customizable ad-hoc heuristics for table detection and cell structure reconstruction based on features of text and ruling lines presented in PDF documents, including the following: horizontal and vertical distances, fonts, the order of appearance of text printing instructions in PDF files and positions of the drawing cursor. Additionally, ad-hoc heuristics can be handcrafted and tuned for various domain-specific PDF documents. The table is subdivided into rows and columns, using the analysis of connected components (text blocks). The proposed method reaches a high recall (0.923) and precision (0.950) on the ICDAR 2013 competition dataset [27].

[4] http://www.icst.pku.edu.cn/cpdp/sjzy/index.htm

3 Table Analysis

The second task (*table analysis*) is to transform a spreadsheet table from an arbitrary to a relational form. This task relies on recovering metadata on the logical structure and content of an arbitrary table, i.e. data items presented in cells, including their functional roles and internal relationships. The recovered semantics enables representing extracted data as relational tables in a spreadsheet-like format (e.g. CSV). In this form, each column of a relational table matches to a category. However, the columns typically remain anonymous, i.e. there is no description of the presented categories. The categories (concepts describing extracted data items) often need to be found out for the purposes of data analysis.

There are several recent studies dealing with the issues of the spreadsheet data extraction and transformation, including the following: layout features [35, 11, 16], "code smells" and formulas [31, 15, 5, 34], programming by examples [6, 59, 32], data models [1, 12, 13], linked open data [45, 71], domain-specific [64, 9, 60] and rule-based architectures [57, 67, 68]. Typically, the related solutions (e.g. [17, 19, 10]) rely on a predefined table structure. They support only a few widespread layout types of tables with typical functional cell regions. This limits their applicability for specific cases.

Our approach consists in using rules for table analysis [53]. The rules map explicit features (layout, style, and text of cells) of an arbitrary table to its implicit semantic relationships (linked functional data items such as entries, labels, and categories). The approach expects that one ruleset provides functional and structural analysis for tables with the same features. Such rulesets can be executed by a rule engine [53, 57] or be translated to executable programs in a general-purpose language [50, 51].

For implementation of our approach, we develop own domain-specific language of table analysis and interpretation rules, CRL [52, 56, 57]. This language determines queries (conditions) and operations (actions) that are necessary to develop programs for spreadsheet data transformation from an arbitrary to relational form. CRL rules expressed as productions map the physical structure of cells to the logical structure of data items. In comparison with general-purpose rule languages (such as Drools[5], Jess[6], RuleML[7]), our language enables expressing rulesets without any instructions for management of the working memory (such as updates of modified facts, or blocks on the rule re-activation). This provides syntactically simplifying declaration of the right-side hand of CRL rules. CRL allows end-users to focus more on the logic of table analysis and interpretation than on the logic of the rule management and execution.

The interpreter of CRL rules provides translating CRL rulesets (declarative programs) to Java source code (imperative programs) [50, 51]. The generated source code is ready for compilation and building of executable programs for domain-specific spreadsheet data extraction and transformation. The novelty of

[5] https://www.drools.org
[6] https://www.jessrules.com
[7] http://ruleml.org

the software platform consist in providing two rule-based ways to implement workflows of spreadsheet data extraction and transformation. In the first case, a ruleset for table analysis and interpretation is expressed in a general-purpose rule language and executed by a JSR-94[8] compatible rule engine. In the second case, our interpreter translates a ruleset expressed in CRL to Java source code that is compiled and executed by using Java platform.

The existing solutions with similar goals (e.g. [17, 19, 10]) typically use predefined table models embedded into their internal algorithms. Unlike them, we define a general-purpose table model that does not restrict layout types. Instead of a generic approach when functions (roles) are associated with cells, in our model functions are determined for data items (entries and labels) originated from cells. The model supports a table layout where one cell contains two or more data items. This allows expressing user-defined layout, style, and text features of arbitrary tables in external rules to support both widespread and specific table types.

4 Table Interpretation

The third task (*table interpretation*) serves to link the extracted tabular data to external vocabularies. Each column and each data item of a relational table should be associated with a concept (class, object, or property) of a general-purpose or domain-specific ontology (Fig. 4). The linked open data (LOD[9]) cloud, including global taxonomies (e.g. DBpedia[10], Wikidata[11], or YAGO[12]) can be utilized as external vocabularies for these purposes. The tabular data enriched by links to external vocabularies are represented in RDF/OWL (Resource Description Framework[13] / Web Ontology Language[14]) formats. The integration of the extracted data with the LOD cloud simplifies the implementation of their further analysis.

The methods intended for the issues of the table interpretation are mainly knowledge-based. They try to bind text in tables with some external concepts, using the following techniques: extraction ontologies [20], data frames [61], knowledge (classes and relations) automatically collected from the Web [63], natural language processing, including the named entity linking [7, 66, 71] and the word embedding [18], as well as various general-purpose ontologies of Linked Open Data [36, 39, 49, 14, 37, 46, 45] or proprietary global taxonomy, ProBase [65]. There are also several studies that propose to use contextual information that surrounds tables [8, 28, 70, 69]. The issues of converting data presented in

[8] https://www.jcp.org/ja/jsr

[9] https://lod-cloud.net

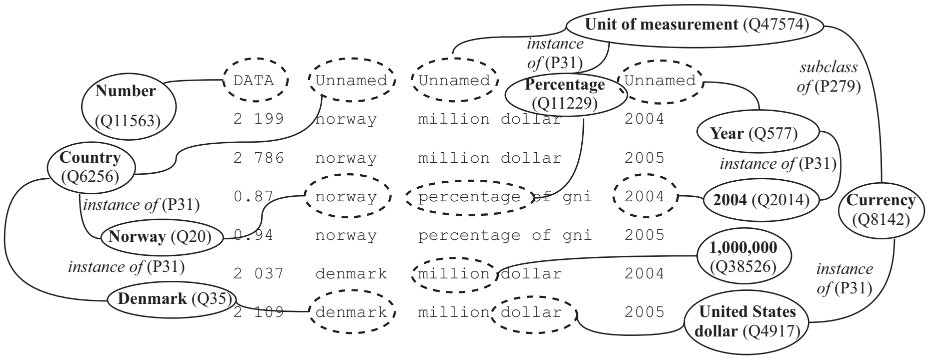
[10] https://wiki.dbpedia.org

[11] https://www.wikidata.org

[12] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago

[13] https://www.w3.org/RDF

[14] https://www.w3.org/OWL



DATA	Unnamed	Unnamed	Unnamed
2 199	norway	million dollar	2004
2 786	norway	million dollar	2005
0.87	norway	percentage of gni	2004
0.94	norway	percentage of gni	2005
2 037	denmark	million dollar	2004
2 109	denmark	million dollar	2005

Fig. 2. A fragment of a table being annotated by using Wikidata's items and properties.

spreadsheets or web tables to RDF/OWL formats are considered in the papers [3, 38, 23, 24, 21].

We suggest an approach to the semantic table interpretation based on combining natural language processing techniques and external vocabularies. First of all, the extracted data items should be separated into two types: numeric and non-numeric via named entity recognition. Then non-numeric data items are linked with concepts (classes, objects, and properties) of an external vocabulary by using the semantic similarity.

To examine this approach we used DBpedia but it can be extended by other vocabularies in future. Our implementation of the approach supports the following functionality: (i) tabular data cleansing and formatting in accordance with DBpedia naming conventions; (ii) creating queries to DBpedia in SPARQL language; (iii) and linking extracted data items of a table with DBpedia classes, objects, and properties.

We developed a prototype of the software tool for generating linked data in RDF/OWL formats from extracted tabular data. It is designed to be a part of an end-to-end process of the table understanding implemented by our software platform. In this environment, it can be applied for the semantic interpretation of tables with an arbitrary layout.

5 Conclusions

The presented approaches can draw up the basis to design the methodology and software for creating systems of data extraction from arbitrary tables contained in weakly structured (e.g. PDF) and semi-structured documents (e.g. spreadsheets). The approaches correspond to the state-of-the-art level studies in the area of information extraction. They rely on the modern techniques of the deep-learning, rule-based and generative programming, linked open data, and table understanding.

Our approach to the table extraction can be quickly adapted to various domain-specific PDF documents (such as financial statements, business credit assessments, material safety data sheets, etc.) with a rich tabular content. Additional ad-hoc heuristics can be handcrafted and tuned for the target domain. This does not require to prepare training datasets that can be a costly process.

The proposed approach to the table analysis involves the rule and generative programming. It includes a principally novel formal language for table analysis and interpretation that should provide expressing table transformation rules. The main advantage of our approach to the semantic table interpretation is that it can be applied for tables with an arbitrary layout. In comparison to our competitors, we support not only widespread layout types of arbitrary tables, but also specific ones.

We expect that the explained principles can be used for designing the software for end-to-end tabular transformation in scientific and industrial data-intensive applications. Particularly, we used them to develop the software for filling up a data warehouse with socio-economic data on Mongolian provinces, for populating the database of the web-based statistical atlas from tabular data of government statistical reports, and for generating ontologies from data of arbitrary tables used in industrial safety inspection.

6 Acknowledgments

This work was financially supported by the Russian Scientific Foundation (Grant No. 18-71-10001).

References

1. Amalfitano, D., Fasolino, A.R., Tramontana, P., De Simone, V., Di Mare, G., Scala, S.: A reverse engineering process for inferring data models from spreadsheet-based information systems: An automotive industrial experience. In: Data Management Technologies and Applications. pp. 136–153 (2015)
2. Arif, S., Shafait, F.: Table detection in document images using foreground and background features. In: Digital Image Computing: Techniques and Applications. pp. 1–8 (2018). https://doi.org/10.1109/DICTA.2018.8615795
3. van Assem, M., Rijgersberg, H., Wigham, M., Top, J.: Converting and annotating quantitative data tables. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) The Semantic Web – ISWC 2010. pp. 16–31 (2010)
4. Bansal, A., Harit, G., Roy, S.D.: Table extraction from document images using fixed point model. In: Indian Conf. on Computer Vision Graphics and Image Processing. pp. 67:1–67:8. ICVGIP '14 (2014). https://doi.org/10.1145/2683483.2683550
5. Barowy, D.W., Berger, E.D., Zorn, B.: ExceLint: Automatically finding spreadsheet formula errors. Proc. ACM Program. Lang. **2**(OOPSLA), 148:1–148:26 (2018). https://doi.org/10.1145/3276518
6. Barowy, D.W., Gulwani, S., Hart, T., Zorn, B.: FlashRelate: Extracting relational data from semi-structured spreadsheets using examples. SIGPLAN Not. **50**(6), 218–228 (2015). https://doi.org/10.1145/2813885.2737952

7. Bhagavatula, C.S., Noraset, T., Downey, D.: Tabel: Entity linking in web tables. In: Arenas, M., Corcho, O., Simperl, E., Strohmaier, M., d'Aquin, M., Srinivas, K., Groth, P., Dumontier, M., Heflin, J., Thirunarayan, K., Thirunarayan, K., Staab, S. (eds.) The Semantic Web - ISWC 2015. pp. 425–441 (2015)
8. Braunschweig, K.: Recovering the Semantics of Tabular Web Data. Ph.D. thesis, Technischen Universität Dresden, Dresden, Germany (2015)
9. Cao, T.D., Manolescu, I., Tannier, X.: Extracting linked data from statistic spreadsheets. In: Proc. Int. Workshop Semantic Big Data. pp. 5:1–5:5 (2017). https://doi.org/10.1145/3066911.3066914
10. Chen, Z.: Information Extraction on Para-Relational Data. Ph.D. thesis, University of Michigan, US (2016)
11. Chen, Z., Dadiomov, S., Wesley, R., Xiao, G., Cory, D., Cafarella, M., Mackinlay, J.: Spreadsheet property detection with rule-assisted active learning. In: Proc. ACM on Conf. on Information and Knowledge Management. pp. 999–1008 (2017). https://doi.org/10.1145/3132847.3132882
12. Cunha, J., Fernandes, J.P., Mendes, J., Saraiva, J.: Embedding, evolution, and validation of model-driven spreadsheets. IEEE Transactions on Software Engineering **41**(3), 241–263 (2015). https://doi.org/10.1109/TSE.2014.2361141
13. Cunha, J., Erwig, M., Mendes, J., Saraiva, J.: Model inference for spreadsheets. Autom Softw Eng **23**(3), 361–392 (2016). https://doi.org/10.1007/s10515-014-0167-x
14. Deng, D., Jiang, Y., Li, G., Li, J., Yu, C.: Scalable column concept determination for web tables using large knowledge bases. Proc. VLDB Endow. **6**(13), 1606–1617 (2013). https://doi.org/10.14778/2536258.2536271
15. Dou, W., Xu, C., Cheung, S.C., Wei, J.: CACheck: Detecting and repairing cell arrays in spreadsheets. IEEE Transactions on Software Engineering **43**(3), 226–251 (2017). https://doi.org/10.1109/TSE.2016.2584059
16. Dou, W., Han, S., Xu, L., Zhang, D., Wei, J.: Expandable group identification in spreadsheets. In: Proc. 33rd ACM/IEEE Int. Conf. on Automated Software Engineering. pp. 498–508 (2018). https://doi.org/10.1145/3238147.3238222
17. Eberius, J., Werner, C., Thiele, M., Braunschweig, K., Dannecker, L., Lehner, W.: DeExcelerator: a framework for extracting relational data from partially structured documents. In: Proc. 22nd ACM Int. Conf. on Information & Knowledge Management. pp. 2477–2480 (2013). https://doi.org/10.1145/2505515.2508210
18. Efthymiou, V., Hassanzadeh, O., Rodriguez-Muro, M., Christophides, V.: Matching web tables with knowledge base entities: From entity lookups to entity embeddings. In: d'Amato, C., Fernandez, M., Tamma, V., Lecue, F., Cudré-Mauroux, P., Sequeda, J., Lange, C., Heflin, J. (eds.) The Semantic Web – ISWC 2017. pp. 260–277 (2017)
19. Embley, D.W., Krishnamoorthy, M.S., Nagy, G., Seth, S.: Converting heterogeneous statistical tables on the web to searchable databases. Int. J. Document Analysis and Recognition **19**(2), 119–138 (2016). https://doi.org/10.1007/s10032-016-0259-1
20. Embley, D., Tao, C., Liddle, S.: Automating the extraction of data from HTML tables with unknown structure. Data Knowl. Eng. **54**(1), 3–28 (2005). https://doi.org/10.1016/j.datak.2004.10.004
21. Ermilov, I., Ngomo, A.C.N.: TAIPAN: Automatic Property Mapping for Tabular Data, pp. 163–179 (2016). https://doi.org/10.1007/978-3-319-49004-5_11
22. Fan, M., Kim, D.S.: Table region detection on large-scale PDF files without labeled data. CoRR **abs/1506.08891** (2015), http://arxiv.org/abs/1506.08891

23. Fiorelli, M., Lorenzetti, T., Pazienza, M.T., Stellato, A., Turbati, A.: Sheet2RDF: a flexible and dynamic spreadsheet import&lifting framework for RDF. In: Proc. 28th Int. Conf. Industrial, Engineering and Other Applications of Applied Intelligent Systems. pp. 131–140 (2015). https://doi.org/10.1007/978-3-319-19066-2_13
24. Galkin, M., Mouromtsev, D., Auer, S.: Identifying web tables: Supporting a neglected type of content on the web. In: Proc. 6th Int. Conf. Knowledge Engineering and Semantic Web. pp. 48–62 (2015). https://doi.org/10.1007/978-3-319-24543-0_4
25. Gao, L., Yi, X., Jiang, Z., Hao, L., Tang, Z.: Icdar2017 competition on page object detection. In: 14th IAPR Int. Conf. on Document Analysis and Recognition. vol. 01, pp. 1417–1422 (2017). https://doi.org/10.1109/ICDAR.2017.231
26. Gilani, A., Qasim, S.R., Malik, I., Shafait, F.: Table detection using deep learning. In: 14th IAPR Int. Conf. on Document Analysis and Recognition. vol. 01, pp. 771–776 (2017). https://doi.org/10.1109/ICDAR.2017.131
27. Göbel, M., Hassan, T., Oro, E., Orsi, G.: Icdar 2013 table competition. In: 12th Int. Conf. on Document Analysis and Recognition. pp. 1449–1453 (2013). https://doi.org/10.1109/ICDAR.2013.292
28. Govindaraju, V., Zhang, C., Ré, C.: Understanding tables in context using standard NLP toolkits. In: Proc. 51st Annual Meeting of the Association for Computational Linguistics. vol. 2: Short Papers, pp. 658–664 (2013)
29. Hao, L., Gao, L., Yi, X., Tang, Z.: A table detection method for pdf documents based on convolutional neural networks. In: 12th IAPR Workshop on Document Analysis Systems. pp. 287–292 (2016). https://doi.org/10.1109/DAS.2016.23
30. He, D., Cohen, S., Price, B., Kifer, D., Giles, C.L.: Multi-scale multi-task fcn for semantic page segmentation and table detection. In: 14th IAPR Int. Conf. on Document Analysis and Recognition. vol. 01, pp. 254–261 (2017). https://doi.org/10.1109/ICDAR.2017.50
31. Hermans, F., Pinzger, M., Deursen, A.: Detecting and refactoring code smells in spreadsheet formulas. Empirical Softw. Engg. **20**(2), 549–575 (2015). https://doi.org/10.1007/s10664-013-9296-2
32. Jin, Z., Anderson, M.R., Cafarella, M., Jagadish, H.V.: Foofah: Transforming data by example. In: Proc. ACM Int. Conf. Management of Data. pp. 683–698 (2017). https://doi.org/10.1145/3035918.3064034
33. Kavasidis, I., Palazzo, S., Spampinato, C., Pino, C., Giordano, D., Giuffrida, D., Messina, P.: A saliency-based convolutional neural network for table and chart detection in digitized documents. CoRR **abs/1804.06236** (2018), http://arxiv.org/abs/1804.06236
34. Koch, P., Hofer, B., Wotawa, F.: On the refinement of spreadsheet smells by means of structure information. Journal of Systems and Software **147**, 64–85 (2019). https://doi.org/10.1016/j.jss.2018.09.092
35. Koci, E., Thiele, M., Romero, O., Lehner, W.: Table identification and reconstruction in spreadsheets. In: Proc. 29th Int. Conf. Advanced Information Systems Engineering. pp. 527–541 (2017). https://doi.org/10.1007/978-3-319-59536-8_33
36. Limaye, G., Sarawagi, S., Chakrabarti, S.: Annotating and searching web tables using entities, types and relationships. Proc. VLDB Endow. **3**(1-2), 1338–1347 (2010). https://doi.org/10.14778/1920841.1921005
37. Muñoz, E., Hogan, A., Mileo, A.: Using linked data to mine RDF from wikipedia's tables. In: Proc. 7th ACM Int. Conf. Web Search and Data Mining. pp. 533–542 (2014). https://doi.org/10.1145/2556195.2556266
38. Mulwad, V., Finin, T., Joshi, A.: A Domain Independent Framework for Extracting Linked Semantic Data from Tables, pp. 16–33 (2012). https://doi.org/10.1007/978-3-642-34213-4_2

39. Mulwad, V., Finin, T., Syed, Z., Joshi, A.: Using linked data to interpret tables. In: Proc. 1st Int. Conf. Consuming Linked Data. vol. 665, pp. 109–120 (2010), http://dl.acm.org/citation.cfm?id=2878947.2878957
40. Perez-Arriaga, M., Estrada, T., Abad-Mota, S.: Tao: System for table detection and extraction from pdf documents (2016), https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS16/paper/view/12916
41. Rashid, S.F., Akmal, A., Adnan, M., Aslam, A.A., Dengel, A.: Table recognition in heterogeneous documents using machine learning. In: 14th IAPR Int. Conf. on Document Analysis and Recognition. vol. 01, pp. 777–782 (2017). https://doi.org/10.1109/ICDAR.2017.132
42. Rastan, R., Paik, H.Y., Shepherd, J.: Texus: A unified framework for extracting and understanding tables in pdf documents. Information Processing & Management **56**(3), 895–918 (2019). https://doi.org/https://doi.org/10.1016/j.ipm.2019.01.008
43. Rastan, R., Paik, H.Y., Shepherd, J., Ryu, S.H., Beheshti, A.: Texus: Table extraction system for pdf documents. In: Wang, J., Cong, G., Chen, J., Qi, J. (eds.) Databases Theory and Applications. pp. 345–349 (2018)
44. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. CoRR **abs/1506.01497** (2015), http://arxiv.org/abs/1506.01497
45. Ritze, D., Bizer, C.: Matching web tables to dbpedia - A feature utility study. In: Proc. 20th Int. Conf. on Extending Database Technology. pp. 210–221 (2017). https://doi.org/10.5441/002/edbt.2017.20
46. Ritze, D., Lehmberg, O., Bizer, C.: Matching HTML tables to DBpedia. In: 5th Int. Conf. on Web Intelligence, Mining and Semantics. pp. 10:1–10:6 (2015). https://doi.org/10.1145/2797115.2797118
47. Schreiber, S., Agne, S., Wolf, I., Dengel, A., Ahmed, S.: Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In: 14th IAPR Int. Conf. on Document Analysis and Recognition. vol. 01, pp. 1162–1167 (2017). https://doi.org/10.1109/ICDAR.2017.192
48. Shahab, A., Shafait, F., Kieninger, T., Dengel, A.: An open approach towards the benchmarking of table structure recognition systems. In: 9th IAPR Int. Workshop on Document Analysis Systems. pp. 113–120 (2010). https://doi.org/10.1145/1815330.1815345
49. Shen, W., Wang, J., Luo, P., Wang, M.: LIEGE: Link entities in web lists with knowledge base. In: 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. pp. 1424–1432 (2012). https://doi.org/10.1145/2339530.2339753
50. Shigarov, A., Khristyuk, V., Mikhailov, A.: Tabbyxl: Software platform for rule-based spreadsheet data extraction and transformation. SoftwareX **10**, 100270 (2019). https://doi.org/https://doi.org/10.1016/j.softx.2019.100270
51. Shigarov, A., Khristyuk, V., Mikhailov, A., Paramonov, V.: Software development for rule-based spreadsheet data extraction and transformation. In: 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics. pp. 1132–1137 (2019). https://doi.org/10.23919/MIPRO.2019.8756829
52. Shigarov, A.: Rule-based table analysis and interpretation. In: Proc. 21st Int. Conf. Information and Software Technologies. pp. 175–186 (2015). https://doi.org/10.1007/978-3-319-24770-0_16
53. Shigarov, A.: Table understanding using a rule engine. Expert Systems with Applications **42**(2), 929–937 (2015). https://doi.org/10.1016/j.eswa.2014.08.045

54. Shigarov, A., Altaev, A., Mikhailov, A., Paramonov, V., Cherkashin, E.: Tabbypdf: Web-based system for pdf table extraction. In: 24th Int. Conf. on Information and Software Technologies. pp. 257–269 (2018)
55. Shigarov, A., Mikhailov, A., Altaev, A.: Configurable table structure recognition in untagged PDF documents. In: Proc. ACM Symposium on Document Engineering. pp. 119–122 (2016). https://doi.org/10.1145/2960811.2967152
56. Shigarov, A., Paramonov, V., Belykh, P., Bondarev, A.: Rule-based canonicalization of arbitrary tables in spreadsheets. In: Proc. 22nd Int. Conf. Information and Software Technologies. pp. 78–91 (2016). https://doi.org/10.1007/978-3-319-46254-7_7
57. Shigarov, A.O., Mikhailov, A.A.: Rule-based spreadsheet data transformation from arbitrary to relational tables. Information Systems **71**, 123–136 (2017). https://doi.org/10.1016/j.is.2017.08.004
58. Siddiqui, S.A., Malik, M.I., Agne, S., Dengel, A., Ahmed, S.: Decnt: Deep deformable cnn for table detection. IEEE Access **6**, 74151–74161 (2018). https://doi.org/10.1109/ACCESS.2018.2880211
59. Singh, R., Gulwani, S.: Transforming spreadsheet data types using examples. SIGPLAN Not. **51**(1), 343–356 (2016). https://doi.org/10.1145/2914770.2837668
60. Swidan, A., Hermans, F.: Semi-automatic extraction of cross-table data from a set of spreadsheets. In: Barbosa, S., Markopoulos, P., Paternò, F., Stumpf, S., Valtolina, S. (eds.) End-User Development. pp. 84–99 (2017)
61. Tijerino, Y., Embley, D., Lonsdale, D., Ding, Y., Nagy, G.: Towards ontology generation from tables. World Wide Web: Internet and Web Information Systems **8**(3), 261–285 (2005). https://doi.org/10.1007/s11280-005-0360-8
62. Tran, D.N., Tran, T.A., Oh, A., Kim, S.H., Na, I.S.: Table detection from document image using vertical arrangement of text blocks. International Journal of Contents **11**(4), 77–85 (2015). https://doi.org/http://dx.doi.org/10.5392/IJoC.2015.11.4.077
63. Venetis, P., Halevy, A., Madhavan, J., Paşca, M., Shen, W., Wu, F., Miao, G., Wu, C.: Recovering semantics of tables on the web. Proc. VLDB Endow. **4**(9), 528–538 (2011). https://doi.org/10.14778/2002938.2002939
64. de Vos, M., Wielemaker, J., Rijgersberg, H., Schreiber, G., Wielinga, B., Top, J.: Combining information on structure and content to automatically annotate natural science spreadsheets. Int. J. Human-Computer Studies **103**, 63–76 (2017). https://doi.org/10.1016/j.ijhcs.2017.02.006
65. Wang, J., Wang, H., Wang, Z., Zhu, K.Q.: Understanding tables on the web. In: Proc. 31st Int. Conf. Conceptual Modeling. pp. 141–155 (2012). https://doi.org/10.1007/978-3-642-34002-4_11
66. Wu, T., Yan, S., Piao, Z., Xu, L., Wang, R., Qi, G.: Entity linking in web tables with multiple linked knowledge bases. In: Li, Y.F., Hu, W., Dong, J.S., Antoniou, G., Wang, Z., Sun, J., Liu, Y. (eds.) Semantic Technology. pp. 239–253 (2016)
67. Yang, S., Guo, J., Wei, R.: Semantic interoperability with heterogeneous information systems on the internet through automatic tabular document exchange. Information Systems **69**, 195–217 (2017). https://doi.org/10.1016/j.is.2016.10.010
68. Yang, S., Wei, R., Shigarov, A.: Semantic interoperability for electronic business through a novel cross-context semantic document exchange approach. In: Proc. ACM Symposium on Doc. Eng. pp. 28:1–28:10 (2018). https://doi.org/10.1145/3209280.3209523
69. Yoshida, M., Matsumoto, K., Kita, K.: Table topic models for hidden unit estimation. In: Proc. 12th Asia Information Retrieval Societies Conference. pp. 302–307 (2016). https://doi.org/10.1007/978-3-319-48051-0_23

70. Zhang, Z.: Towards Efficient and Effective Semantic Table Interpretation, pp. 487–502 (2014). https://doi.org/10.1007/978-3-319-11964-9_31

71. Zhang, Z.: Effective and efficient semantic table interpretation using TableMiner$^+$. Semantic Web **8**(6), 921–957 (2017). https://doi.org/10.3233/SW-160242