data

MDPI

Data Descriptor

Classification of Actual Sensor Network Deployments in Research Studies from 2013 to 2017

Janis Judvaitis [†], Artis Mednis, Valters Abolins, Ansis Skadins, Didzis Lapsa, Raimonds Rava, Maksims Ivanovs and Krisjanis Nesenbergs [*,†]

Institute of Electronics and Computer Science, 14 Dzerbenes St., LV-1006 Riga, Latvia; janis.judvaitis@edi.lv (J.J.); artis.mednis@edi.lv (A.M.); valters.abolins@edi.lv (V.A.); ansis.skadins@edi.lv (A.S.); didzis.lapsa@edi.lv (D.L.); raimonds.rava@edi.lv (R.R.); maksims.ivanovs@edi.lv (M.I.)

[*] Correspondence: krisjanis.nesenbergs@edi.lv

[†] These authors contributed equally to this work.

Received: 17 August 2020; Accepted: 24 September 2020; Published: 30 September 2020

Abstract: Technologies, such as Wireless Sensor Networks (WSN) and Internet of Things (IoT), have captured the imagination of researchers, businesses, and general public, due to breakthroughs in embedded system development, sensing technologies, and ubiquitous connectivity in recent years. That resulted in the emergence of an enormous, difficult-to-navigate body of work related to WSN and IoT. In an ongoing research effort to highlight trends and developments in these technologies and to see whether they are actually deployed rather than subjects of theoretical research with presumed potential use cases, we gathered and codified a dataset of scientific publications from a five-year period from 2013 to 2017 involving actual sensor network deployments, which will serve as a basis for future in-depth analysis of the field. In the first iteration, 15,010 potentially relevant articles were identified in SCOPUS and Web of Science databases; after two iterations, 3059 actual sensor network deployments were extracted from those articles and classified in a consistent way according to different categories, such as type of nodes, field of application, communication types, etc. We publish the resulting dataset with the intent that its further analysis may identify prospective research fields and future trends in WSN and IoT.

Dataset: https://doi.org/10.5281/zenodo.4048214

Dataset License: CC0

Keywords: WSN; IoT; deployment; review

1. Summary

As we are heading into the 21st century, digitalization trends in transportation [1,2], in-house logistics [3], education [4,5], agriculture [6], banking [7,8], and other fields are providing new and engaging ways for technology to improve our daily lives. This naturally leads to the emergence of applications of Wireless Sensor Networks (WSN) and Internet of Things (IoT) in large number of different domains. The WSN and IoT popularity is growing rapidly and, according to Grand View Research, the Narrow Band IoT (NB-IoT) market size will reach more than $6 billion by 2025 [9]. Yet, the majority of researchers still use simulation tools to validate their theories [10] rather than deploy actual devices; as a consequence, it is unclear to what extent the vast majority of the available WSN/IoT devices are actually used instead of theorized as being applicable and what design choices drive the selection of devices.

The aim of this work was to provide a comprehensive high level mapping of actual WSN and IoT deployments used by the research community to serve as a foundation for future in-depth analysis of related trends from the five-year period from 2013 to 2017. The presented dataset can be further used for various statistical and contextual analysis, as well as further extended to cover a broader time frame. As the complete marked data set is available, together with intermediate collection results, the authenticity of the data can be verified.

Altogether, 15,010 data articles were identified as potential candidates, from which after two iterations of screening 3059 actual sensor network deployments were extracted and codified according to multiple categories, as described in the next sections.

The data acquisition, analysis, and validation took around two years for a team of 12 volunteer researchers, of which eight provided significant value.

The rest of the document is structured, as follows—Section 2 describes the data set as such, Section 3 discusses methods that are used in acquiring the data as well as data validation and quality, and finally Section 4 contains some practical data usage notes.

2. Data Description

The dataset contains data files that result from the data acquisition process as shown in Table 1 and described below in detail. The files are in one of three formats:

- .bib—BibTeX format containing entries representing published articles;
- .json—JSON format containing structured human readible data object entries; and,
- .txt—text files containing TAB delimited tabular data with a header row.

Table 1. Data files in the dataset.

ID	File Name	Number of Data Entries	Description
(A)	0_Merged_Full_Collection_15010.bib	15,010	Identified candidate articles from database keyword search as a BibTeX file
(B)	1_Screened_4915.bib	4915	Candidate articles left after screening as a BibTeX file
(C)	1_Screening_statistics.txt	12	Screening statistics per person as Tab delimited text file
(D)	1_Screening_timeline.txt	17	Screening timeline per week of screening as Tab delimited text file
(E)	2_Eligibility_statistics.txt	12	Eligibility check statistics per person as Tab delimited text file
(F)	2_Eligible_3017.json	3017	Candidate articles left after eligibility check as a JSON file
(G)	2_Ineligible_1898.json	1898	Articles excluded in the eligibility check and reason for exclusion as a JSON file
(H)	2_Mistaken_as_ineligible_47.json	47	Articles mistakenly excluded in the eligibility check and re-included during validation as a JSON file
(I)	3_Eligibility_and_extraction_timeline.txt	35	Eligibility check and data extraction weekly timeline as Tab delimited text file
(J)	**3_Extracted_data_3059.json**	**3059**	**Main dataset—Deployments identified and codified data extracted as a JSON file**
(K)	3_Extraction_statistics.txt	12	Data extraction statistics per person as Tab delimited text file
(L)	3_Mistaken_as_eligible_15.json	15	Articles mistakenly included in the eligibility step and excluded during data extraction as a JSON file
(M)	Timeline.txt	14	Overall timeline of dataset building process as a Tab delimited text file
(N)	README.md	-	Readme file with short description of the dataset
(O)	Notebooks	-	Folder containing Jupyter notebook files for easier data visualisation with processing examples

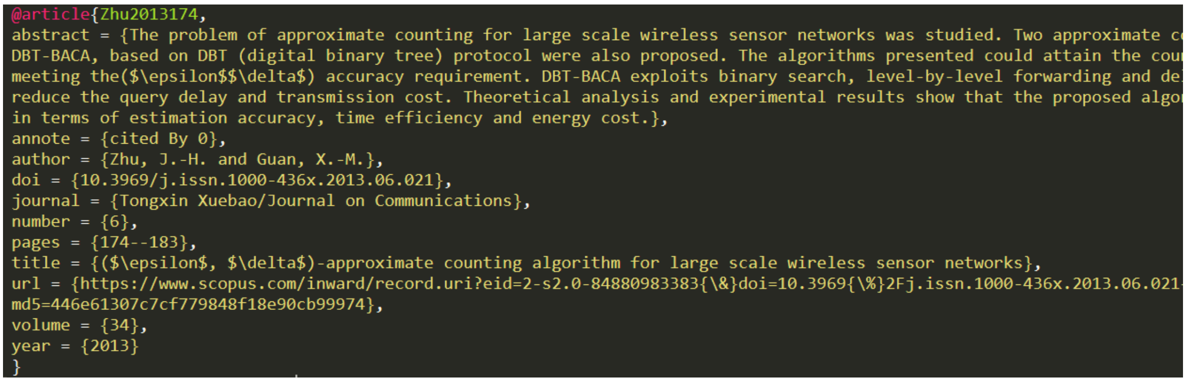
In the subsections below, the technical description of data entries with possible data types and values are described in detail. Verbatim data values in this description will be formatted, like `this`.

For the eager reader interested in the main resulting dataset, please refer to dataset (**J**) on page 7.

2.1. (A)—Identified Candidate Articles

This file contains 15,010 BibTeX entries, which have the following types: `@article` (7137), `@book` (74), `@conference` (1861), `@incollection` (67) and `@inproceedings` (5871).

Each entry in the file starts on a new line, and can continue over multiple lines. An example entry can be seen in Figure 1. The basic structure of an entry is `@type{id,metadata}`, where `type` is name of the document type e.g., article or book, `id` is a unique string in the document identifying that specific entry and `metadata` is a list of comma separated key/value pairs describing the entry. Not all entries contain the same metadata entries, but most have the following: `abstract`, `author`, `doi`, `title`, and `year`. Additionally, depending on the entry type, additional metadata, like `page`, `volume`, or `url`, could be present.



```
@article{Zhu2013174,
abstract = {The problem of approximate counting for large scale wireless sensor networks was studied. Two approximate c
DBT-BACA, based on DBT (digital binary tree) protocol were also proposed. The algorithms presented could attain the cou
meeting the($\epsilon$$\delta$) accuracy requirement. DBT-BACA exploits binary search, level-by-level forwarding and de
reduce the query delay and transmission cost. Theoretical analysis and experimental results show that the proposed algo
in terms of estimation accuracy, time efficiency and energy cost.},
annote = {cited By 0},
author = {Zhu, J.-H. and Guan, X.-M.},
doi = {10.3969/j.issn.1000-436x.2013.06.021},
journal = {Tongxin Xuebao/Journal on Communications},
number = {6},
pages = {174--183},
title = {($\epsilon$, $\delta$)-approximate counting algorithm for large scale wireless sensor networks},
url = {https://www.scopus.com/inward/record.uri?eid=2-s2.0-84880983383{\&}doi=10.3969{\%}2Fj.issn.1000-436x.2013.06.021
md5=446e61307c7cf779848f18e90cb99974},
volume = {34},
year = {2013}
}
```

Figure 1. Data entry example in dataset (A) and (B).

2.2. *(B)—Screened Candidate Articles*

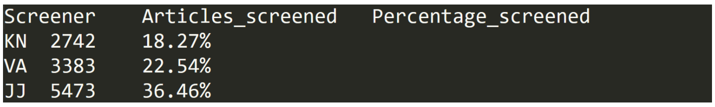
This file contains 4915 BibTeX entries of the same format, as described in previous section, thus the related entry format is also shown in Figure 1. These entries represent candidate articles left from the (A) dataset after first step of screening and the file contains the following entry types: `@article` (2385, 33% left ater screening), `@book` (12, 16% left), `@conference` (569, 31% left), `@incollection` (12, 18% left), and `@inproceedings` (1937, 33% left).

2.3. *(C)—Screening Statistics*

This file is formatted as a table in a TAB delimited text file. It has 12 entries, each pertaining to one of the 12 volunteer researchers involved in the screening process.

Each row contains and entry (see Figure 2 for entry examples) that has the following three headers/columns with corresponding data types:

1. `Screener`—two letter code uniquely identifying each of the researchers. Example of data in column: `KN`;
2. `Articles_screened`—number of articles processed by the corresponding researcher in the screening step. This is an integer value in range from `0` to `5473`;
3. `Percentage_screened`—the percentage of the total number of articles in dataset (A) that were processed by the researcher in the screening step. This number is formed as percentage value rounded to two decimal places and has values from `0.00%` to `36.46%`.

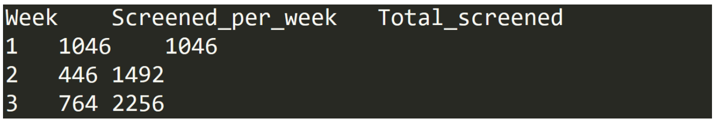


```
Screener    Articles_screened    Percentage_screened
KN   2742    18.27%
VA   3383    22.54%
JJ   5473    36.46%
```

Figure 2. Data entry examples in dataset (C).

2.4. *(D)—Screening Timeline*

The file is formatted as a table in a TAB delimited text file. It has 17 entries, which each represent one of the 17 weeks, during which the screening process took place (for example entries, see Figure 3).



```
Week    Screened_per_week   Total_screened
1   1046    1046
2   446 1492
3   764 2256
```

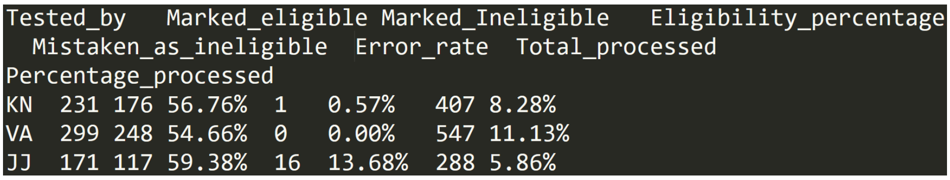
Figure 3. Data entry examples in dataset (D).

Each row has the following three headers/columns with corresponding data types:

1. `Week` —number of the week in screening process, represented by an integer value in range from `1 to 17;`
2. `Screened_per_week` —number of articles processed during the specific screening week by all researchers involved. This is an integer value in range from `50 to 2068;`
3. `Total_screened` —cumulative number of articles processed up to and including the specific screening week by all researchers involved. This is an integer value in range from `1046` to `15,010.`

2.5. *(E)—Eligibility Statistics*

This file is formatted as a table in a TAB delimited text file. It has 12 entries, where each pertain to one of the 12 volunteer researchers involved in the eligibility checking process. Figure 4 shows example entries.



```
Tested_by   Marked_eligible Marked_Ineligible   Eligibility_percentage
  Mistaken_as_ineligible  Error_rate  Total_processed
Percentage_processed
KN  231 176 56.76%  1   0.57%   407 8.28%
VA  299 248 54.66%  0   0.00%   547 11.13%
JJ  171 117 59.38%  16  13.68%  288 5.86%
```

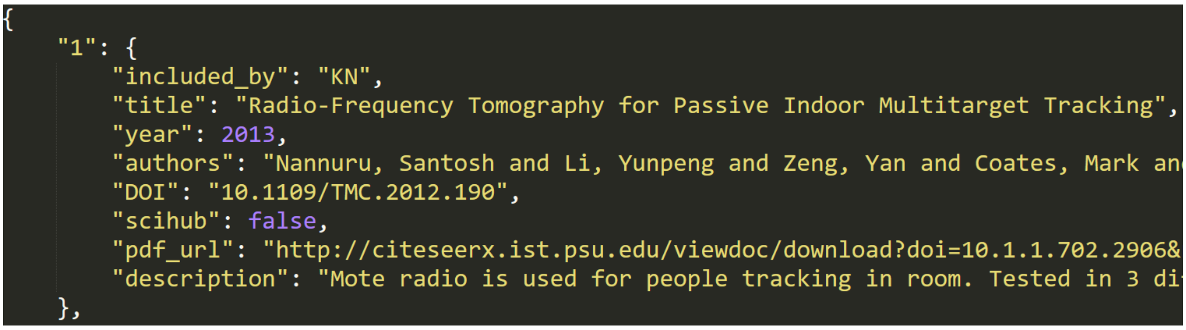
Figure 4. Data entry examples in dataset (E).

Each row has the following eight headers/columns with corresponding data types:

1. `Tested_by` —two letter code uniquely identifying each of the researchers. Example of data in column: `KN;`
2. `Marked_eligible` —number of articles processed and marked as eligible by the corresponding researcher in the eligibility checking step. This is an integer value in range from `1 to 708;`
3. `Marked_Ineligible` —number of articles processed and marked as ineligible by the corresponding researcher in the eligibility checking step. This is an integer value in range from `0 to 475;`
4. `Eligibility_percentage` —the percentage of the total number of articles checked by the researcher in eligibility checking step that were marked as eligible. This number is formed as percentage value rounded to 2 decimal places and has values from `54.66% to 100.00%.`
5. `Mistaken_as_ineligible` —number of articles mistakenly marked as ineligible by the corresponding researcher in the eligibility checking step. This is an integer value in range from `0 to 20;`
6. `Error_rate` —the percentage of the total number of articles checked by the researcher in eligibility checking step that were mistakenly marked as ineligible. This number is formed as percentage value rounded to two decimal places and it has values from `0.00% to 33.33%.` Additionally, one value is `NaN` or "not a number" representing value resulting from division by zero;
7. `Total_processed` —number of articles processed by the corresponding researcher in the eligibility checking step. This is an integer value in range from `1 to 1183;`
8. `Percentage_processed` —the percentage of the total number of articles in dataset (B) that were processed by the researcher in the eligibility checking step. This number is formed as percentage value rounded to two decimal places and has values from `0.02% to 24.07%.`

2.6. (F)—Candidate Articles Marked as Eligible

This file contains a JSON data object with 3017 entries, each representing a single article that is marked as eligible in the eligibility checking step. Figure 5 shows an example entry .



```
{
    "1": {
        "included_by": "KN",
        "title": "Radio-Frequency Tomography for Passive Indoor Multitarget Tracking",
        "year": 2013,
        "authors": "Nannuru, Santosh and Li, Yunpeng and Zeng, Yan and Coates, Mark an
        "DOI": "10.1109/TMC.2012.190",
        "scihub": false,
        "pdf_url": "http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.702.2906&
        "description": "Mote radio is used for people tracking in room. Tested in 3 di
    },
```

Figure 5. Data entry example in dataset (F).

The object is structured, as follows: `{entry1, entry2, ..., entry3017}`. Each of the entries have the following structure: `id:{key1:value1, ..., key8:value8}`, where `id` is a unique string identifier of the entry (e.g., `"42"`) and each key/value pair represents one of eight metadata entries from the Table 2 below. In some cases, where a specific metadata value was not available for an entry, the value can also be `null`.

Table 2. Metadata format in dataset (F).

Key	Value Type	Description	Example/Possible Values
`"included_by"`	string	Two character unique identifier of researcher who marked this article as eligible.	`"KN"`
`"title"`	string	Text string containing full title of the eligible article	`"Some title"`
`"year"`	integer	Year, when the article was published	`from 2013 to 2017`
`"authors"`	string	Text string containing list of authors of the article in BibTeX format	`"Author, A and Author, B"`
`"DOI"`	string	Text string representing Digital Object Identifier (DOI) of the publication	`"10.1109/IE.2015.30"`
`"scihub"`	boolean	Wether article was publicly available (false) or had to be acquired through other channels (true)	`true or false`
`"pdf_url"`	string	Contains URL of the publicly available article if it exists	`"https://someurl.com"`
`"description"`	string	Text string containing a short description of the contents of the article	`"Self calibrating WSN"`

2.7. (G)—Candidate Articles Marked as Ineligible

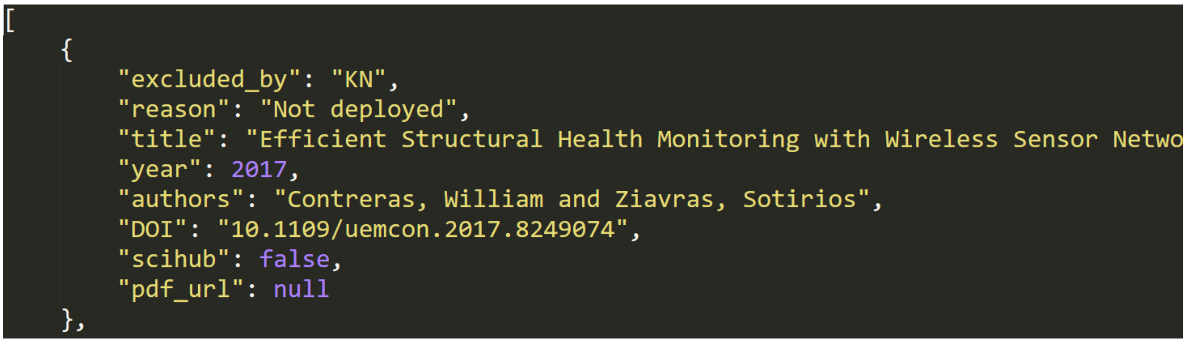
This file contains a list of 1898 JSON data objects, which each represent a single article marked as ineligible in the eligibility checking step.

The list is structured, as follows: `[entry1, entry2, ..., entry3017]`. Each of the entries have the following structure: `{key1:value1, ..., key8:value8}`, where each key/value pair represents one of eight metadata entries from Table 3 below. In some cases, where a specific metadata value was not available for an entry, the value can also be `null`. Figure 6 shows an example entry .

Because only articles that describe actual physical deployment of sensor network devices (more than one and networked) were included, several groups of articles were excluded, as ilustrated by the `"reason"` metadata field, which can take one of the following values (number of matching entries in the dataset in brackets):

- `"Article not available"` (438 entries)—we were not able to access full text of the article;
- `"Theoretical"` (160 entries)—the article described theoretical aspects not practical deployment;
- `"Not deployed"` (293 entries)—no deployment was described even though device might be developed;

- `"Article not English"` (88 entries)—article not available in English language;
- `"Simulation"` (485 entries)—experiment was simulated thus not using actual deployment;
- `"No network"` (183 entries)—non-networked devices (usually data loggers) or a single device deployed;
- `"Review"` (23 entries)—a review article of other deployment articles, excluded to avoid duplication; and,
- `"Other"` (166 entries)—some other reason for exclusion—usually not related to sensor networks at all.



```
[
    {
        "excluded_by": "KN",
        "reason": "Not deployed",
        "title": "Efficient Structural Health Monitoring with Wireless Sensor Netwo
        "year": 2017,
        "authors": "Contreras, William and Ziavras, Sotirios",
        "DOI": "10.1109/uemcon.2017.8249074",
        "scihub": false,
        "pdf_url": null
    },
```

Figure 6. Data entry example in dataset (G).

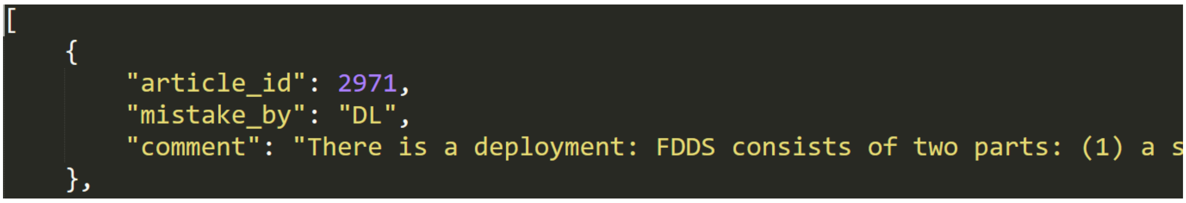
Table 3. Metadata format in dataset (G).

Key	Value Type	Description	Example/Possible Values
`"excluded_by"`	string	Two character unique identifier of researcher who marked this article as ineligible.	`"KN"`
`"reason"`	string	Text string containing a short reason for exclusion of the contents of the article	See explanation above
`"title"`	string	Text string containing full title of the eligible article	`"Some title"`
`"year"`	integer	Year, when the article was published	from 2013 to 2017
`"authors"`	string	Text string containing list of authors of the article in BibTeX format	`"Author, A and Author, B"`
`"DOI"`	string	Text string representing Digital Object Identifier (DOI) of the publication	`"10.1109/IE.2015.30"`
`"scihub"`	boolean	Wether article was publicly available (false) or had to be acquired through other channels (true)	`true` or `false`
`"pdf_url"`	string	Contains URL of the publicly available article if it exists	`"https://someurl.com"`

2.8. (H)—Candidate Articles Mistaken as Ineligible

This file contains a list of 47 JSON data objects, which each represent a single article marked as ineligible by mistake, even though it was actually eligible, during the eligibility checking step.

The list is structured, as follows: `[entry1, entry2, ..., entry47]`. Each of the entries have the following structure: `{key1:value1, key2:value2, key3:value3}`, where each key/value pair represents one of three metadata entries from the Table 4, below. Figure 7 shows an example entry.



```
[
    {
        "article_id": 2971,
        "mistake_by": "DL",
        "comment": "There is a deployment: FDDS consists of two parts: (1) a s
    },
```

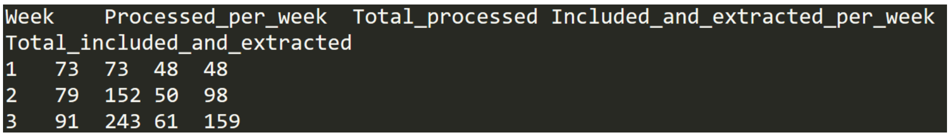
Figure 7. Data entry example in dataset (H).

Table 4. Metadata format in dataset (H).

Key	Value Type	Description	Example/Possible Values
`"article_id"`	integer	An integer identifier of the article, matching the id field in dataset (F)	from `2971` to `3017`
`"mistake_by"`	string	Two character unique identifier of researcher who made the mistake in marking the article as ineligible	`"KN"`
`"comment"`	string	Text string containing short explanation why the article should be included	`"Some comment"`

2.9. (I)—Timeline of Eligibility Check and Data Extraction Phase

This file is formatted as a table in a TAB delimited text file. It has 35 entries, which each represent one of the 35 weeks during which the eligiblity checking and data extraction phase took place. Figure 8 shows an example entry.



```
Week    Processed_per_week  Total_processed Included_and_extracted_per_week
Total_included_and_extracted
1   73  73  48  48
2   79  152 50  98
3   91  243 61  159
```

Figure 8. Data entry example in dataset (I).

Each row has the following five headers/columns with coresponding data types:

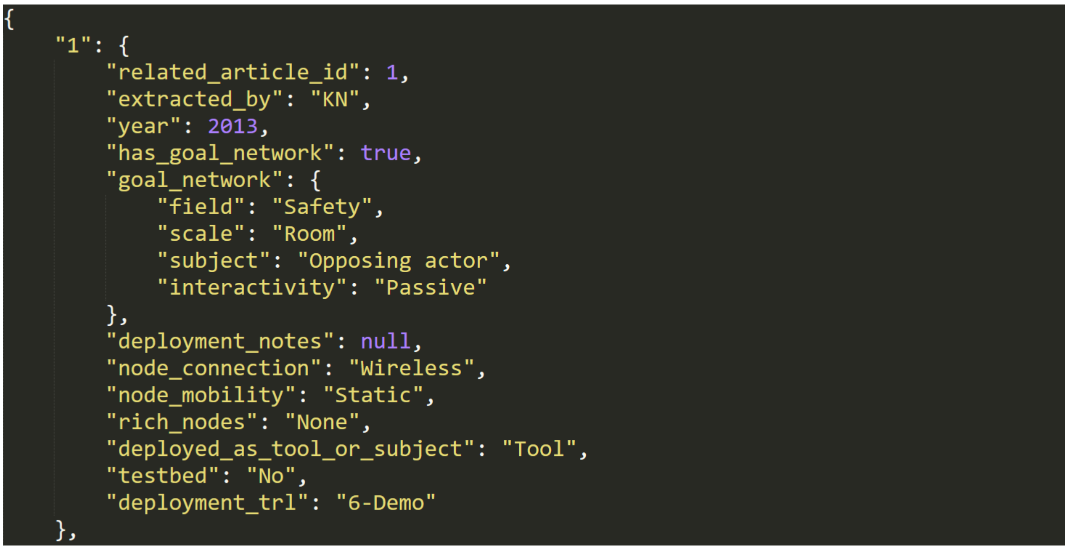
1. `Week` —the number of week for which statistics is given. This is an integer value in range from `1` to `35`;
2. `Processed_per_week` —the number of articles processed per week in the eligiblity checking and data extraction phase. This is an integer value in range from `56` to `302`;
3. `Total_processed` —the cumulative number of articles processed up to and including that week. This is an integer value in range from `73` to `4915`;
4. `Included_and_extracted_per_week` —the number of articles included and actually used for data extraction per week. This is an integer value in range from `33` to `186`;
5. `Total_included_and_extracted` —the cumulative number of articles included and actually used for data extraction up to and including that week. This is an integer value in the range from `48` to `2970`.

2.10. (J)—Extracted Codified Data

This file contains a JSON data object with 3059 entries, which each represent a single deployment from the previously identified articles and containing extracted codified data that are related to this deployment.

The object is structured, as follows: `{entry1, entry2, ..., entry3059}`. Each of the entries have the following structure: `id:{key1:value1, ..., key12:value12}` where `id` is a unique string identifier of the entry (e.g., `"42"`) and each key/value pair represents one of 12 metadata entries from the Table 5 below. In some cases, where a specific metadata value was not available for an entry, the value can also be `null`. Figure 9 shows an example entry.

In the context of these data, a device is considered to be a “rich” device instead of ordianry sensor network device, if it is an interactive computer like system with some multimedia capabilities, e.g., smartphone, personal computer, Raspberry PI, etc.



```
{
    "1": {
        "related_article_id": 1,
        "extracted_by": "KN",
        "year": 2013,
        "has_goal_network": true,
        "goal_network": {
            "field": "Safety",
            "scale": "Room",
            "subject": "Opposing actor",
            "interactivity": "Passive"
        },
        "deployment_notes": null,
        "node_connection": "Wireless",
        "node_mobility": "Static",
        "rich_nodes": "None",
        "deployed_as_tool_or_subject": "Tool",
        "testbed": "No",
        "deployment_trl": "6-Demo"
    },
```

Figure 9. Data entry example in dataset (J).

Table 5. Metadata format in dataset (J).

Key	Value Type	Description	Example/Possible Values
"related_article_id"	integer	Identifier number of the article from dataset (F) in which the specific deployment was described	from 1 to 3017
"extracted_by"	string	Two character unique identifier of the researcher who extracted data about this deployment	"KN"
"year"	integer	Year, when the article containing the deployment was published	from 2013 to 2017
"has_goal_network"	boolean	Wether deployment is made with goal application in mind (true, 1825 entries) instead of just technology focused (false, 1234 entries)	true or false
"goal_network"	data object	Contains data object describing the goal network if it exists	See below
"deployment_notes"	string	Text string containing optional comments by extracting researcher on the deployment	"Vague description..."
"node_connection"	string	Type of connection between sensor nodes—wireless (2860 entries), wired (94 entries), hybrid using both types (89 entries), or not defined (16 entries)	"Wireless", "Wired", "Hybrid", or null
"node_mobility"	string	Mobility type of sensor nodes—static only (2286 entries), mobile only (580 entries), mixed—some static and some mobile (140 entries), or not defined (53 entries)	"Static", "Mobile", "Mixed", or null
"rich_nodes"	string	Which sensor nodes are "rich" devices—none of the nodes are rich (2374 entries), only base station nodes are rich (416 entries), all nodes are rich (226 entries), mixed—some simple and some rich nodes (27 entries), or not defined (16 entries)	"None", "Base_stations", "All", "Mixed", or null
"deployed_as_tool_or_subject"	string	Whether the deployment in the article is used as a tool in the research described (1618 entries), or is the subject of the research itself (1441 entries)	"Tool" or "Subject"
"testbed"	string	Whether a sensor network testbed is used for the described sensor network deployment (478 entries), isn't used (2516 entries) or whether the sensor network itself is part of a testbed (65 entries)	"Used", "No" or "Part of"
"deployment_trl"	string	The Technology Readiness Level of the deployment: 3-bench tested concept (103 entries), 4-validated in laboratory (682 entries), 5-tested in artificial environment/testbed (888 entries), 6-demonstrated on close-to-real environment (479 entries), 7-demonstrated in real environment, or 8-final system in real environment (81 entry)	"3-Bench", "4-Lab", "5-Test", "6-Demo", "7-Target", or "8-Final"

In addition to the overall description of the sensor network deployment itself, such as type of connection of sensor nodes and technology readiness level of the deployment, as described in the article, an additional group of metadata was extracted related to the potential future goal network that the reserach is building towards. Although a major part of the deployments is driven by technology development (1234 entries) not application and don't have such a goal network, for those deployments that have some practical application in mind (1825 entries), the following metadata object is stored under the key `"goal_network": {key1:value1, ..., key4:value4}`. In this object for each deployment, four keys with these possible values and number of entries in dataset are provided:

- `"field"`—The target field of application with one of the following values:
  1. `"Health & wellbeing"`, including patient, frail, and elderly monitoring systems, sports performance, and general health and wellbeing of both body and mind (349 entries);
  2. `"Education"` including systems meant for educational purpouses and serious games (10 entries);
  3. `"Entertainment"` including computer games, AR/VR systems, broadcasting, sporting and public events, gambling and other entertainment (17 entries);
  4. `"Safety"` including anti-theft, security, privacy enhancing, reliability improving, emergency response and military applications and tracking people and objects for these applications (163 entries);
  5. `"Agriculture"` including systems related to farming, crop growing, farm and domesticated animal monitoring, precision agriculture (229 entries);
  6. `"Environment"` monitoring of environment both in wild life and city, including weather, pollution, wild life, forest fires, aquatic life, volcanic activity, flooding, earthquakes etc. (297 entries);
  7. `"Communications"` general communications like power lines, water and gas pipes, energy consumption monitoring, internet, telephony, radio etc. (51 entries);
  8. `"Transport"` inlcuding intelligent transport systems (ITS) smart mobility, logistics and goods tracking, smart road infrastructure etc. (123 entries);
  9. `"Infrastructure"` general infrastructure, such as tunnels, bridges, dams, ports, smart homes and buildings etc. (413 entries);
  10. `"Industry"` anything related to industrial processes, production and business in general like coal mine monitoring, production automation, quality control, process monitoring etc. (143 entries);
  11. `"Research"` not related to other fields, but to support future research—better resaerch tools and protocols, testbeds etc. (20 entries); and,
  12. `"Multiple"` the deployed network will serve multiple of the previously described fields (10 entries).
- `"scale"`—The target deployment scale of the sensor network with one of the following values (from smallest to larges):
  1. `"Single actor"` including such single entities as a person (e.g., body area network), animal, vehicle, or robot (345 entries);
  2. `"Room"` include such relatively small territories as rooms, garages, small yards (131 entries);
  3. `"Building"` include larger areas with separate zones, like houses, private gardens, shops, hospitals (530 entries);
  4. `"Property"` include even larger zones capable of containing multiple buildings, like city blocks, farms, small private forest or orchard (447 entries);
  5. `"Region"` include areas of city or self-government scale like a rural area, forest, lake, river, city or suburbs (317 entries);
  6. `"Country"` include objects of scale relative to countries, like national road grid, large agricultural or forest areas, smaller seas (27 entries);
  7. `"Global"` include networks of scale not limited to a single country, such as oceans, jungle or space (24 entries); and,
  8. `null`—no scale information of target deployment provided or it is not clearly defined (four entries).

- "subject"—The main target subject meant to be monitored by the goal network with one of the following values:

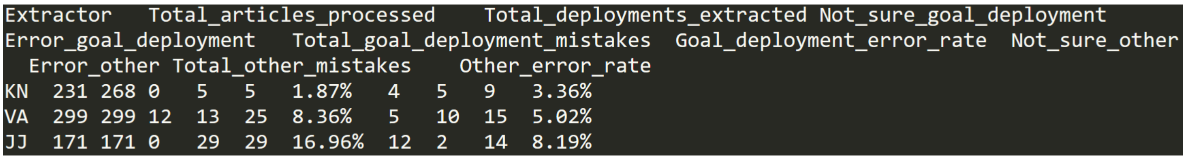
  1. "Environment" includes all types of environmental phenomena, like weather, forests, bodies of water, habitats, etc. (728 entries);
  2. "Equipment" includes all types of inanimate objects, including industrial equipment, buildings, vehicles or robots as systems not actors in environment, dams, walls etc. (498 entries);
  3. "Opposing actor" include all types of actors in environment, which do not want to be monitored, thus including security and spying applications, tracking and monitoring of perpetrators or military opponents, pest control etc. (126 entries);
  4. "Friendly actor" includes actors that do not mind to be tracked or monitored for some purpose, like domestic or wild animals (tagging), elderly or frail, people in general if compliant (456 entries);
  5. "SELF" includes cases where the sensor network monitors itself—location of nodes, communication quality etc. (one entry); and,
  6. "Mixed"—this includes target deployments with multiple subjects from the previously stated values (16 entries).

- "interactivity"—The interactivity of the goal sensor network with the following values:

  1. "Passive" includes passive monitoring nodes and data gathering for decision making outside the system or for general statistics purposes (1448 entries);
  2. "Interactive" includes sensor networks providing some kind of feedback, control or interactivity within the loop or confines of the system, like automated irrigation systems, real time alarms etc. (375 entries); and,
  3. null no specific interactivity of target deployment is provided or clearly defined in the article (two entries).

2.11. (K)—Statistics of Extraction Process

This file is formatted as a table in a TAB delimited text file. It has 12 entries, each pertaining to one of the 12 volunteer researchers that are involved in the data extraction process. Example entries shown in Figure 10.



```
Extractor    Total_articles_processed     Total_deployments_extracted Not_sure_goal_deployment
Error_goal_deployment    Total_goal_deployment_mistakes  Goal_deployment_error_rate  Not_sure_other
  Error_other Total_other_mistakes     Other_error_rate
KN  231 268 0   5   5   1.87%   4   5   9   3.36%
VA  299 299 12  13  25  8.36%   5   10  15  5.02%
JJ  171 171 0   29  29  16.96%  12  2   14  8.19%
```

Figure 10. Data entry examples in dataset (K).

Each row has the following 11 headers/columns with coresponding data types:

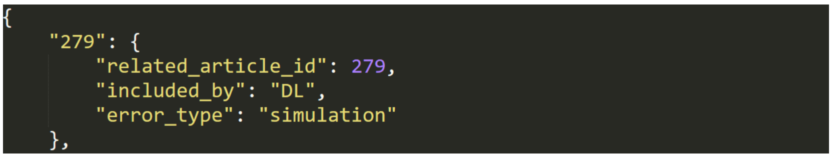
1. Extractor—two letter code uniquely identifying each of the researchers. Example of data in column: KN;
2. Total_articles_processed—number of articles processed by the corresponding researcher in the data extraction step. This is an integer value in range from 1 to 708;
3. Total_deployments_extracted—number of actual sensor network deployments extracted from these articles by the corresponding researcher in the data extraction step. This is an integer value in range from 1 to 708;

4. `Not_sure_goal_deployment`—number of articles in which the extractor was not sure about the goal deployment of the sensor network and required peer input to get the final value. This is an integer value in range from `0` to `12`;
5. `Error_goal_deployment`—number of articles in which the extractor mistakenly marked a wrong goal deployment value, which was later corrected in validation stage. This is an integer value in range from `0` to `79`;
6. `Total_goal_deployment_mistakes`—sum of two previous values representing the total amount of errors related to the goal deployment made by the specific extractor. This is an integer value in range from `0` to `87`;
7. `Goal_deployment_error_rate`—the percentage of the total number of deployments processed by the researcher in data extraction stage that contained some sort of error related to goal deployment data extraction. This number is formed as percentage value rounded to 2 decimal places and has values from `0.00%` to `30.00%`;
8. `Not_sure_other`—number of articles in which the extractor was not sure about the some other metadata value not related to goal deployment and required peer input to get the final value. This is an integer value in range from `0` to `60`;
9. `Error_other`—number of articles in which the extractor mistakenly marked a wrong metadata value not related to goal deployment, which was later corrected in validation stage. This is an integer value in range from `0` to `10`;
10. `Total_other_mistakes`—sum of two previous values representing the total amount of errors not related to the goal deployment made by the specific extractor. This is an integer value in range from `0` to `66`; and,
11. `Other_error_rate`—the percentage of the total number of deployments processed by the researcher in data extraction stage that contained some sort of error not related to goal deployment data extraction. This number is formed as percentage value rounded to two decimal places and has values from `0.00%` to `50.00%`.

2.12. *(L)—Candidate Articles Mistaken as Eligible*

This file contains JSON data object with 15 entries, each representing a single article that is marked as eligible by mistake, even though it was actually ineligible, discovered during the data extraction step.

The object is structured, as follows: `{entry1, entry2, ..., entry15}`. Each of the entries have the following structure: `id:{key1:value1, key2:value2, key3:value3}`, where `id` is a unique string identifier of the article and each key/value pair represents one of 3 metadata entries from the Table 6 below. Figure 11 shows an example entry.



```
{
    "279": {
        "related_article_id": 279,
        "included_by": "DL",
        "error_type": "simulation"
    },
```

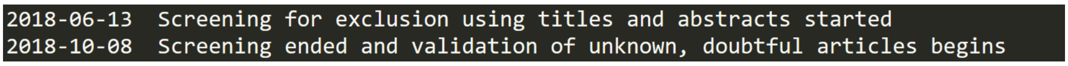
Figure 11. Data entry example in dataset (L).

Table 6. Metadata format in dataset (L).

Key	Value Type	Description	Example/Possible Values
`"related_article_id"`	integer	An integer identifier of the article, matching the id field in dataset (F)	from `279` to `2676`
`"included_by"`	string	Two character unique identifier of researcher who made the mistake in marking the article as eligible	`"KN"`
`"error_type"`	string	Text string describing the reason for exclusion of the mistakenly included deployment, including simulation (3 entries), use of previously existing data (5 entries), no network between devices (4 entries) and sensor network not actually deployed (3 entries)	`"simulation"`, `"existing data"`, `"no network"` or `"no deployment"`

2.13. (M)—Overall Timeline of Dataset Creation

This file is formatted as a table in a TAB delimited text file. It has 14 entries, each pertaining to a milestone date in the progress of dataset creation and has no column headers. Figure 12 shows example entries.



```
2018-06-13  Screening for exclusion using titles and abstracts started
2018-10-08  Screening ended and validation of unknown, doubtful articles begins
```

Figure 12. Data entry examples in dataset (M).

Each row has the following 2 columns with coresponding data types:

1. Date in format of `yyyy-mm-dd` with values in the range from `2018-06-12` to `2020-07-02`; and,
2. Milestone event description in the form of a text string.

2.14. (N)—Readme File

This file contains a short human readable description of the data in this dataset in the form of a Markdown document.

2.15. (O)—Notebook Folder

In this folder, several Jupyter notebook files are stored for easy loading of and access to the data files. These contain example Python 3 code for opening the files and extracting the data within.

3. Methods

To acquire this dataset, the scope of the problem was first defined, as follows: to gather and codify all scientific peer reviewed publications describing original practical sensor network deployments from a five-year period from 2013 to 2017. The scope was narrowed for practical purposes, as follows:

- only publications in English language were considered;
- only publications that could be accessed by the research team without use of additional funds were considered;
- to be considered a network, the deployment had to have at least two actually deployed sensor devices;
- devices did not have to be wireless, to be considered sensor network—also wired, acoustic, or other networks were considered;
- only research doing the deployment themselves was considered—no use of ready datasets from other deployments was included;
- no simulated experiments were included;
- the timeframe was selected as 2013–2017, because the data acquisition was started in the middle of 2018, and only full years were chosen for comparability; and,
- to avoid duplicates only original deployments were included instead of review articles.

Based on this scope, a systematic literature review methodology was devised and followed consisting of the following steps (note that in dataset, the files related to these steps are enumerated starting from 0 not 1):

1. Candidate article acquisition
2. Screening (exclusion)
3. Screening (inclusion/eligibility)
4. Codification and data extraction
5. Verification

3.1. Candidate Article Acquisition

Due to their popularity and wide access in the institutions represented by the authors, two main indexing databases were selected for querying articles: SCOPUS and Web of Science.

For each of these databases, a query with the same information based on the scope defined above was prepared:

- **SCOPUS:** KEY (sensor network OR sensor networks) AND TITLE-ABS-KEY (test* OR experiment* OR deploy*) AND NOT TITLE-ABS-KEY (review) AND NOT TITLE-ABS-KEY (simulat*) AND (LIMIT-TO (PUBYEAR,2017) OR LIMIT-TO (PUBYEAR,2016) OR LIMIT-TO (PUBYEAR,2015) OR LIMIT-TO (PUBYEAR,2014) OR LIMIT-TO (PUBYEAR,2013))
- **Web of Science:** TS = ("sensor network" OR "sensor networks") AND TS = (test* OR experiment* OR deploy*) NOT TI ="review" NOT TS = simulat* **with additional parameters:** Indexes = SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan = 2013–2017

The querying was done on **12 June 2018** and it yielded the following results:

- **SCOPUS:** 11,536 total articles identified of which 4814 were not found in Web of Science database;
- **Web of Science:** 10,204 total articles identified of which 3636 were not found in SCOPUS database;
- After checking for duplicates **15,010 unique candidate articles** were identified of which 6560 articles were found in both databases. Duplicates were checked both automatically while using features that were provided by Mendeley software and manually by title/author/year combination.

The resulting dataset was saved as BibTeX file (see dataset (A)) and imported in Mendeley software for collaborative screening for exclusion.

3.2. First Screening Iteration-Exclusion

During this stage, the research team was instructed to exclude articles conservatively—only exclude those that definitely match the exclusion citeria and leave all others for a more thorough examination in the next stage.

The exclusion criteria was defined, as follows: *The article does not feature a real life deployment of a sensor network or is not in English language.*

The team of volunteers participating in the screening process were all provided access to a shared Mendeley group with the dataset of 15,010 articles and given the following instructions:

1. spend no more than 10 min on a single article;
2. only look at article title and abstract for exclusion;
3. when processing an article mark it as "read" (a gray/green circle mark in Mendelay);
4. if the article matches exclusion citeria move onto the next article;
5. otherwise, include it for the next stage by marking it with favourite/star icon in Mendeley;

6. regularly synchronize progress and follow randomized article slots based on alphabetic order of article titles to avoid collisions of multiple reviewers; and,
7. in the case of doubt, articles could be tagged for second opinion by another reviewer.

One researcher took lead of it and re-evaluated the first 100 articles processed by all other researchers, and discussed any differences or problems, in order to ensure consistent understanding of the exclusion process. Weekly discussions on progress, problematic articles, etc. were held.

The screening for exclusion took place from 13 June 2018 until 8 October 2018 with the weekly progress that is shown in Figure 13. Subsequently, validation stage started, during which the randomized sample or articles was double checked by other researchers and 142 articles identified as requiring second opinion were discussed and marked appropriately. The validation of screening/exclusion phase ended on 3 November 2018 with 4915 articles left for the next stage (thus 10,095 articles were excluded in this phase).

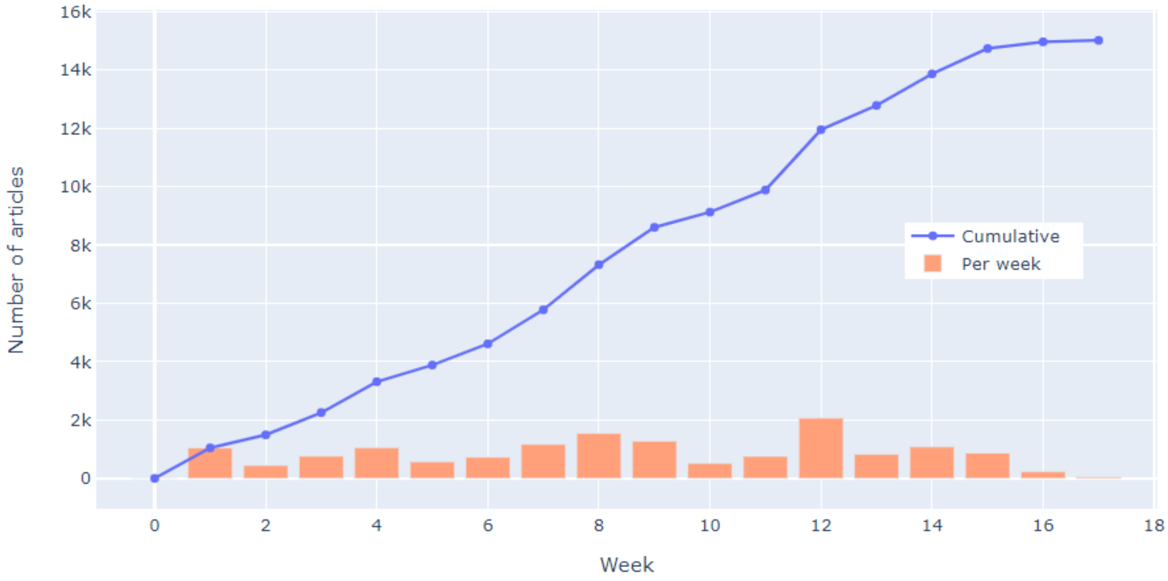


Figure 13. Weekly progress of first screening phase.

3.3. *Second Screening Iteration-Inclusion/Eligibility*

After the first stage of screening, the second screening iteration phase started. This phase required opening and reading the full text of the articles, thus, for time conservation, it was done in parallel with the next phase—codification and data extraction (see next section).

First, from 4 November 2018 till 13 Janaury 2019, an instruction for full text eligibility validation was developed together with data codification and data extraction methodology. An online spreadsheet was developed with the 4915 articles from the previous screening phase, with columns for the required data as dropboxes.

Subsequently, from 14 Janaury 2019 till 15 September 2019 data inclusion/eligibility and codified data extraction stage took place—the weekly progress can be seen in Figure 14. The main steps in this stage for all researchers involved were, as follows:

1. mark the row corresponding to the selected article with unique identifier of the researcher, so that no one else accidentally takes the same article for analysis;
2. locate the full text of the article—if it is not available in English language from any source (indexing pages, preprint publishing pages, author pages, Researchgate, Sci-hub, general google search, etc.), then exclude the article from data extraction, otherwise move to the next step;

3. read the article to identify any sensor network deployments in it. If there are no deployments, then the article must be excluded. If there are several deployments, insert new lines in the table, thus describing each deployment separately;
4. do not include any articles that should have been excluded in the previous stage (review articles, articles without actual deployments or using old data from previous deployments, or even deployments with single sensor device or multiple devices, which have no sensors or network between them;
5. for each included row, leave a comment on which/how many actual sensor network deployments are there—these deployment rows in the spreadhseet table are then filled by the same researcher as part of the next phase (see next Section).

During the second screening stage, 2970 articles were first included and codified. Subsequently, on 17 September 2019 a verification phase of excluded articles was begun, and involved both randomized reviews, as well as multiple reviews of any article marked as uncertain by the original researcher. After this phase ended on 2 December 2019, an additional 47 articles were found in the mistakenly excluded article list and included, thus leading to 3017 total articles eligible for extraction.

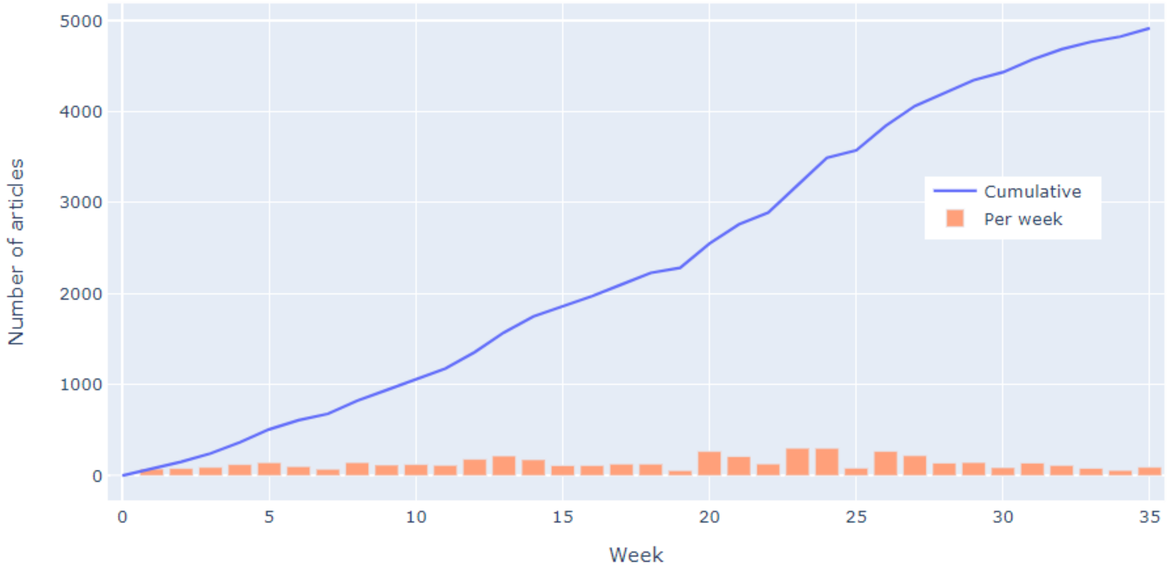


Figure 14. Weekly progress of second screening phase.

From the excluded 4915 − 3017 = 1898 articles, the reasons for exclusion from most frequent to least frequent were: (1) article describes simulation not actual deployment—485 articles; (2) article full text not available—438 articles; (3) sensor network only described, but not actually deployed—293 articles; (4) only separate sensor devices with no network/local data logging—183 articles; (5) theorethical article with no practical experiments—160 articles; (6) article not available in English—88 articles; (7) article uses existing data gathered from a previous deployment or public data set—62 articles; and, (8) Article is a review article of other deployments—23 articles. Additionally, 166 articles were excluded due to other reasons, that didn't correspond to one of the above mentioned categories (e.g., nothing to do with sensor devices or disqualified due to multiple categories).

Until 5 Janaury 2020, all of the deployments in these articles were identified and codified and a thoroguh validation phase of codified data was carried out during the process in which 15 articles were identified as mistakenly included for codification leaving only 3002 articles.

The total number of identified sensor network deployments in these articles was 3059.

3.4. Data Codification and Extraction

For all of the 3059 deployments the researchers involved had to extract two codified groups of data:

1. details on the actual sensor network deployment described in the article; and,
2. if exists—the goal deployment towards which this research is aimed in the future.

The specific codification values are described in detail in the data description of dataset (J), as shown in Section 2.10. In addition to these values, all of the researchers were allowed to provide `null` value if the article did not mention or describe the specific value of interest and `OTHER` value if the researcher did not think that the value could fit in any previously defined category. Additionally on every field the researchers could leave comments asking for second opinion or leaving discussion points about the codification system.

As with the exclusion stage, the data extraction stage also contained coordination between the researchers involved—the first 10 codification efforts by each of the researchers were double checked by one researcher, so that everyone had a common understanding. All of the questions and unclear values were discussed weekly for clarifications.

Finally the codified data was verified—all of the comments were manually processed, outlier values, `null` values, and `OTHER` values were double checked by other researchers, in order to verify that something was not missed by the original reader of the article. Additionally, random validation of codified entries occured.

The errors during validation were labeled and counted for each of the researchers involved (as can be seen in datasets (E), (H), (K), and (L). The deployments that were checked by researchers who were outliers (with low amount of articles processed or high amount of specific errors) were re-checked by other researchers.

Finally, on 29 May 2020, the dataset was completed and preparation started for publishing the data set. Data set was cleaned up, formatted, and submitted to an open access Git repository on 2 July 2020. Afterwards, the text of this publication was prepared together with Jupyter Notebook examples on use of these data.

3.5. Data Quality

In addition to random validation and checking for errors, as described in the previous steps, additional checks on the data set were done to ensure quality of the data.

First, the number of article candidates from each year were compared to see if there is a bias for specific years (e.g., older articles). Each year the number of articles was around the mean 3002, with deviation of less than 4.5%.

Subsequently, the screening phase results were analyzed to test for bias related to year. In all years, 30% to 35% of articles survived the first screening/exclusion phase, with no observable bias towards any particular year (see Figure 15).

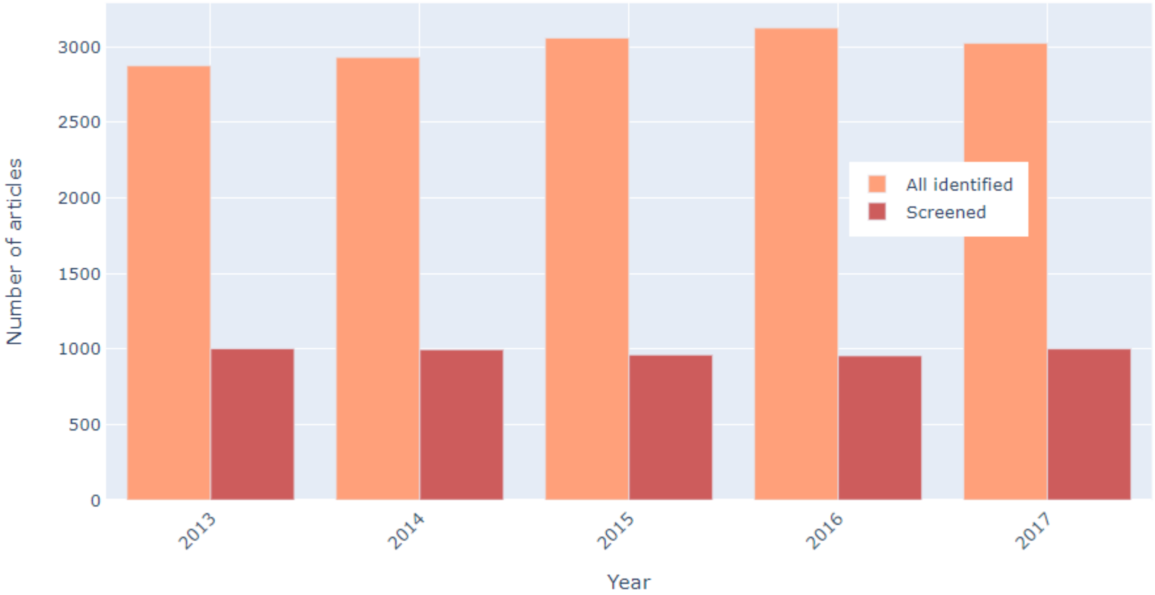


Figure 15. Number of articles initially identified per year and corresponding number of articles included in first screening phase.

The approximate 1/3 inclusion rate also held true for the three most represented categories of articles: `@article` with 33.42%, `@conference` with 30.57% and `@inproceedings` with 32.99% inclusion rates. The two less represented groups `@book` and `@incollection` each had less than 75 instances in the first dataset and, thus, even though their inclusion rate differed from the expected (16.22% and 17.91%, respectively) this is most likely due to the small number of articles in these categories not an inherent bias towards them in the screening process.

Another potential source of bias is the differences in researchers doing the screening, so all of the involved researchers were analyzed. Most had a similar approval ratio of articles (articles marked eligible over all articles processed) and similar low error percentage from total articles processed. As seen in Figure 16, there are three main outliers: `RZ` who has 100% approval ratio, which is due to the fact that this researcher only processed one article in this stage, `SO` whose approval ratio is closer to 70% instead of 60%, like others, which is also due to the low number of articles processed (less than 40), and finally `RB`, who had around 15% error rate in comparison to other researchers who had error rate below 5%. This is also due to the low number of processed articles (7). All other researchers in this phase processed several hundreds of articles and their statistics and error rates were very similar, showing that the efforts to reduce bias that were introduced by individuals were successful.

Overall, wherever a potential source for bias was detected due to a low number of articles being processed by a researcher, their work was re-validated by at least one other researcher to guarantee high data quality.

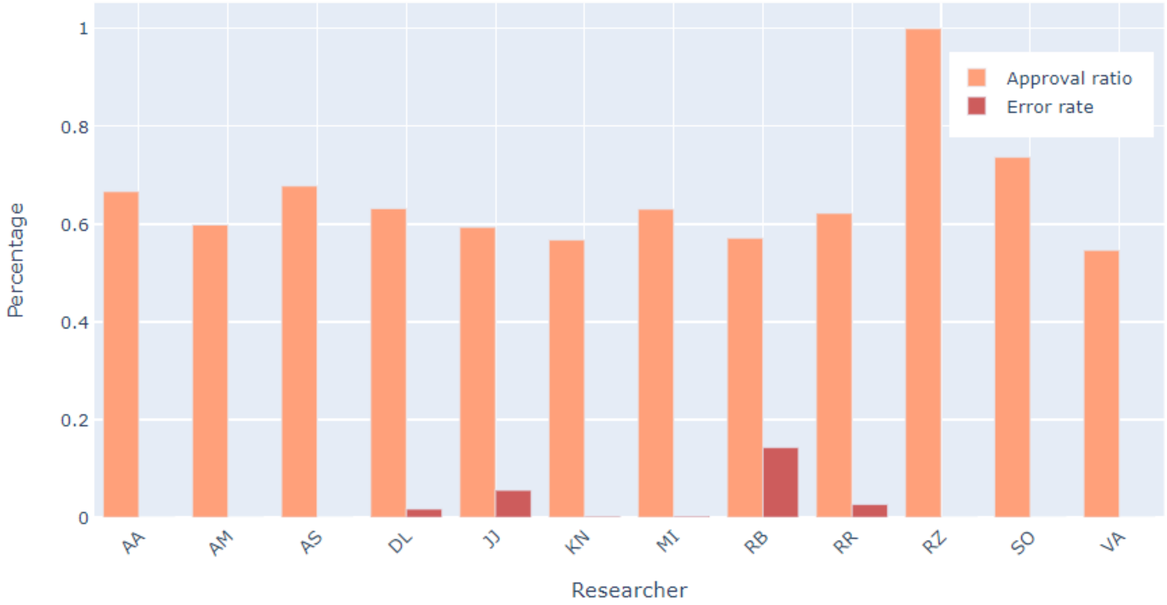


Figure 16. For each researcher—the ratio of articles marked as eligible and their detected error percentage in eligibility screening phase.

4. User Notes

The data set was primarily meant for easy processing while using programming tools, such as Python/Jupyter Notebooks, thus it is machine readible first.

The data is made freely accessible to everybody, although we would appreciate credit if at all possible. To the best of our knowledge, this is the only data set of its kind and currently only covers years 2013 to 2017.

The data is published as a frozen mirror at `https://doi.org/10.5281/zenodo.4048214`.
For user convenience live version can be accessed as a Git repository:

```
git clone http://git.edi.lv/CPS_Lab/sn_deployment_mapping_review
```

In this way, you will get all of the files described in Section 2.

For examples on loading and processing this data using Python, you can access the folder `Notebooks` where Jupyter notebook files with examples on data exploration are stored.

The data structure and examples promote the easy expandability of the dataset—for example, researchers interested in the impact or citation count of the articles containing identified deployments, can use Python libraries such as `scholarly` (for Google Scholar), `wos` (for Web of Science), or `pyscopus` (for SCOPUS) to automatically acquire this additional information—see example in notebook `Explore_extraction_step.ipynb`.

Author Contributions: Conceptualization, J.J. and K.N.; methodology, K.N.; software, K.N.; validation, J.J., A.M., V.A. and K.N.; formal analysis, K.N.; investigation, J.J., A.M., V.A., A.S., D.L., R.R., M.I. and K.N.; data curation, K.N.; writing–original draft preparation, J.J.; writing–review and editing, K.N., M.I. and R.R.; visualization, K.N.; supervision, K.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research is co-financed by ERDF funds under the project No. 1.1.1.1/18/A/183 “iTrEMP: Intelligent transport and emergency management platform”.

Acknowledgments: In addition to the authors of this article also these people provided their work to the acquisition and processing of this data: Rihards Balass, Reinholds Zviedris, Armands Ancans and Sandra Ose.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

WSN	Wireless Sensor Network
IoT	Internet of Things
TAB	Tabulator character
JSON	JavaScript Object Notation

References

1. Noussan, M.; Hafner, M.; Tagliapietra, S. Digitalization Trends. In *The Future of Transport Between Digitalization and Decarbonization*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 51–70.
2. Noussan, M.; Tagliapietra, S. The effect of digitalization in the energy consumption of passenger transport: An analysis of future scenarios for Europe. *J. Clean. Prod.* **2020**, *258*, 120926. [CrossRef]
3. Winkler, H.; Zinsmeister, L. Trends in digitalization of intralogistics and the critical success factors of its implementation. *Braz. J. Oper. Prod. Manag.* **2019**, *16*, 537–549. [CrossRef]
4. Dorofeeva, A.A.; Nyurenberger, L.B. Trends in digitalization of education and training for industry 4.0 in the Russian Federation. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2019; Volume 537, p. 042070.
5. Li, F.; Yang, J.; Wang, J.; Li, S.; Zheng, L. Integration of digitization trends in learning factories. *Procedia Manuf.* **2019**, *31*, 343–348. [CrossRef]
6. Kosareva, O.A.; Eliseev, M.N.; Cheglov, V.P.; Stolyarova, A.N.; Aleksina, S.B. Global trends of digitalization of agriculture as the basis of innovative development of the agro-industrial complex of Russia. *Eurasian J. Biosci.* **2019**, *13*, 1675–1681.
7. Rodin, B.; Ganiev, R.; Orazov, S. «Fintech» in digitalization of banking services. In Proceedings of the International Scientific and Practical Conference on Digital Economy (ISCDE 2019), Chelyabinsk, Russia, 7–8 November 2019; Atlantis Press: Paris, France, 2019.
8. Evdokimova, Y.; Shinkareva, O.; Bondarenko, A. Digital banks: Development trends. In Proceedings of the 2nd International Scientific Conference on New Industrialization: Global, National, Regional Dimension (SICNI 2018), Ekaterinburg, Russia, 4–5 December 2018; Atlantis Press: Paris, France, 2019.
9. NB-IoT Market Size Worth $6.02 Billion by 2025. Available online: https://www.bloomberg.com/press-releases/2019-07-23/nb-iot-market-size-worth-6-02-billion-by-2025-cagr-34-9-grand-view-research-inc (accessed on 16 June 2020).
10. Lima, L.E.; Kimura, B.Y.L.; Rosset, V. Experimental environments for the internet of things: A review. *IEEE Sens. J.* **2019**, *19*, 3203–3211. [CrossRef]