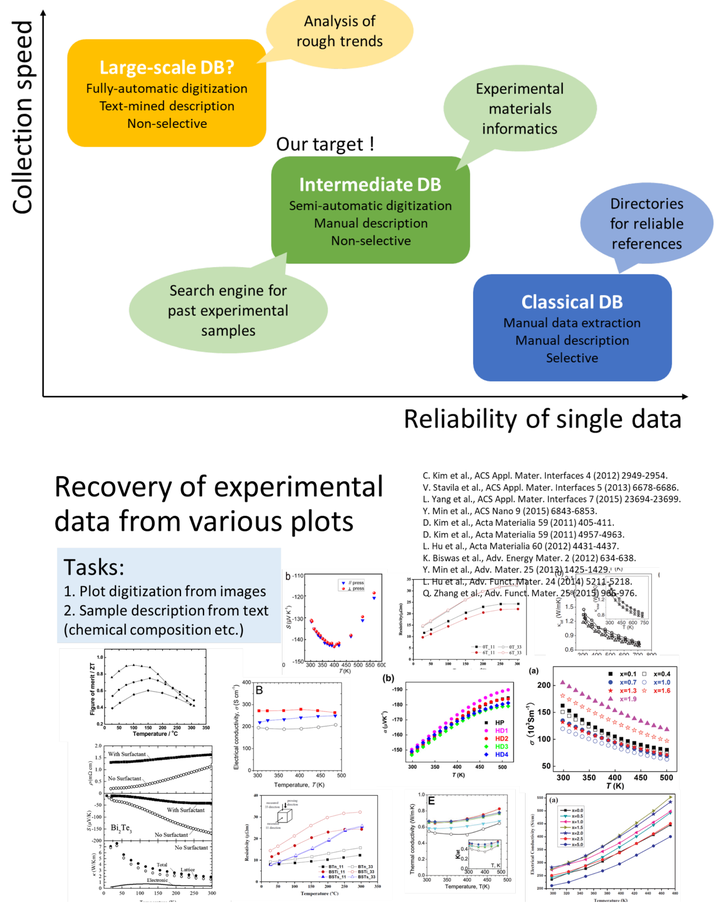
Starrydata: a plot mining web system for literature data

Yukari Katsura[1,2], Masaya Kumagai[3,4], Mitsunori Kaneshige[5], Yuki Ando[2,3], Sakiko Gunji[1,2], Yoji Imai[2,3], Takushi Kodani[1,2], Riku Sato[1,2], Kaoru Kimura[1], Koji Tsuda[1,2,3]

Graduate School of Frontier Sciences, the University of Tokyo[1], NIMS-MI[2]I [2], RIKEN-AIP [3], SAKURA Internet Inc.[4], X-Ability Co.,Ltd.[5]

Plot mining from published papers on materials science



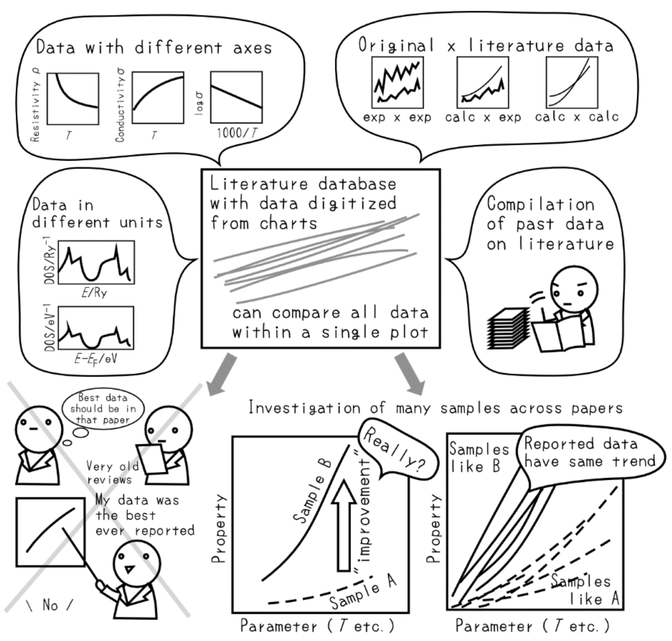
Recovery of experimental data from various plots

Tasks:
1. Plot digitization from images
2. Sample description from text (chemical composition etc.)

C. Kim et al., ACS Appl. Mater. Interfaces 4 (2012) 2949-2954.
V. Stavila et al., ACS Appl. Mater. Interfaces 5 (2013) 6678-6686.
L. Yang et al., ACS Appl. Mater. Interfaces 7 (2015) 23694-23699.
Y. Min et al., ACS Nano 9 (2015) 6843-6853.
D. Kim et al., Acta Materialia 59 (2011) 405-411.
D. Kim et al., Acta Materialia 59 (2011) 4957-4963.
L. Hu et al., Acta Materialia 60 (2012) 4431-4437.
K. Biswas et al., Adv. Energy Mater. 2 (2012) 634-638.
Y. Min et al., Adv. Mater. 25 (2013) 1425-1429.
L. Hu et al., Adv. Funct. Mater. 24 (2014) 5211-5218.
Q. Zhang et al., Adv. Funct. Mater. 25 (2015) 966-976.

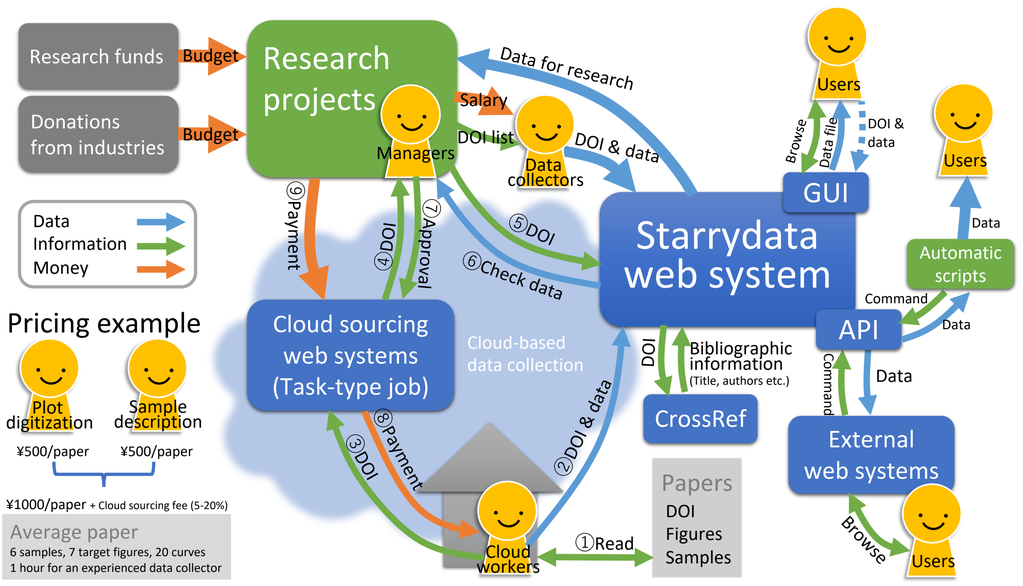
Our goals

1. Make every published data accessible
2. Make research efficient by data science



Our approach

1. Develop an efficient web system for data collection
2. Design an economic system (community) for data collection
3. Share the data freely for worldwide researchers

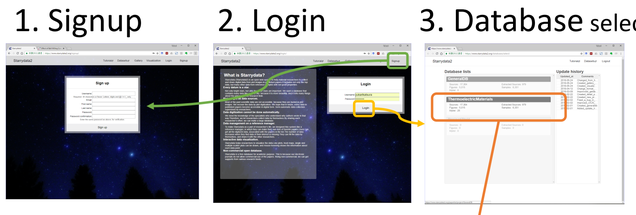


Starrydata web system

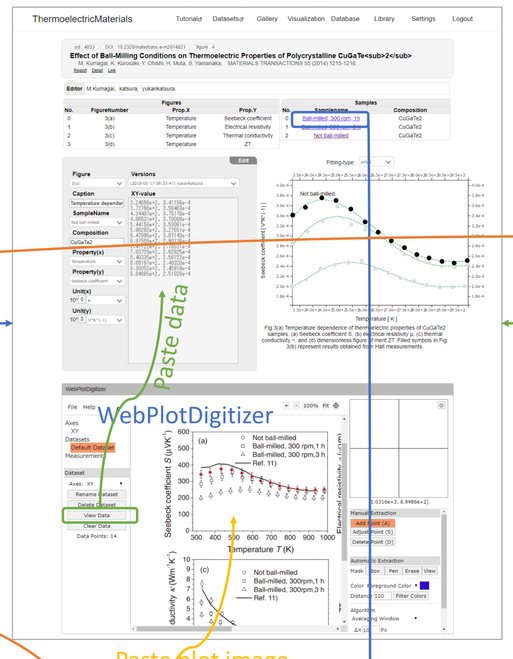
We developed this reference-manager-like web system to share numerical data extracted from plot images written in JavaScript and Python, by using Django, MongoDB, VEGA.js and WebPlotDigitizer (https://automeris.io/WebPlotDigitizer).

Free access
https://www.starrydata2.org

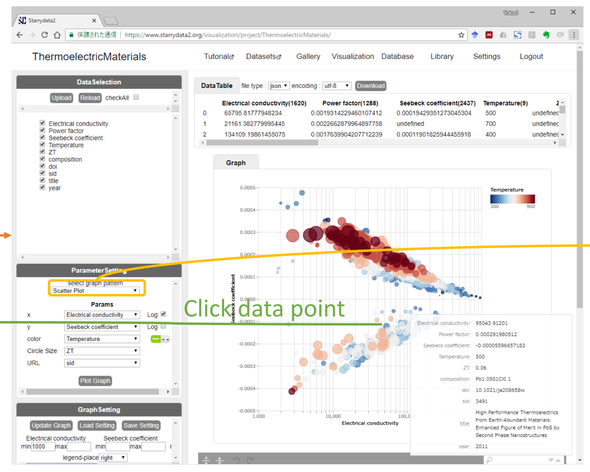
1. Signup 2. Login



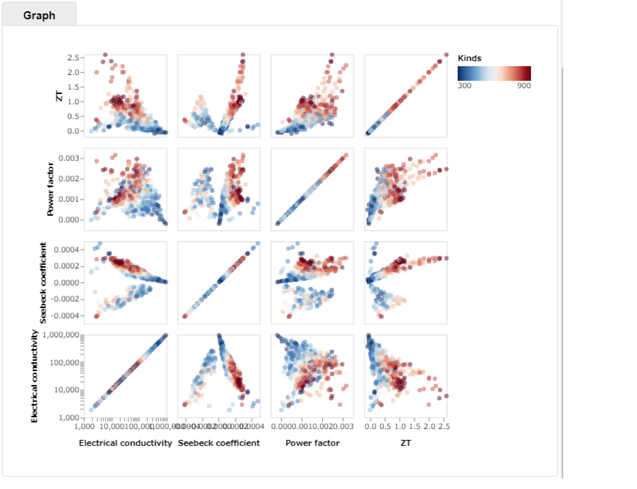
3. Database selection 5. Data upload/editing interface



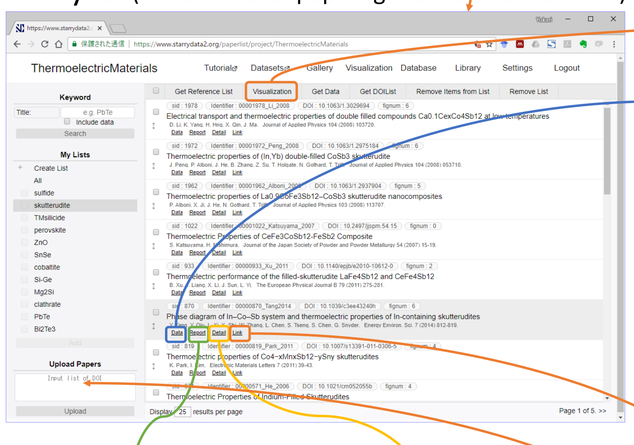
9. Visualization of all data in the mylist



Click data point



4. Mylist (custom lists of papers generated from DOI list)

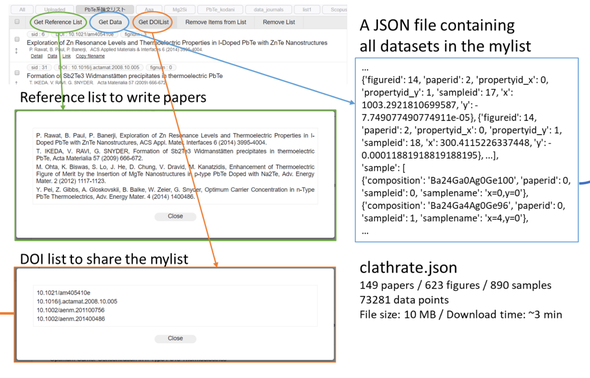


Paste data

WebPlotDigitizer

Paste plot image

10. Data export

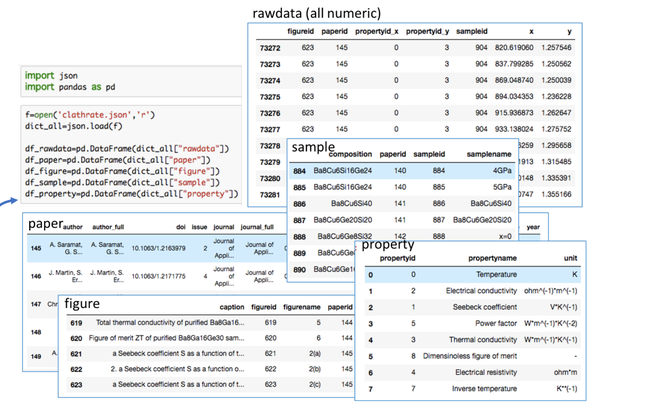


A JSON file containing all datasets in the mylist


```
...
{'figureid': 14, 'paperid': 2, 'propertyid_x': 0, 'propertyid_y': 1, 'sampleid': 17, 'x': 1003.2921810699587, 'y': -7.749077490774911e-05}, {'figureid': 14, 'paperid': 2, 'propertyid_x': 0, 'propertyid_y': 1, 'sampleid': 18, 'x': 300.4115226337448, 'y': -0.0001188191818191188195}, ...],
'sample': [
{'composition': 'Ba24Ga0Ag0Ge100', 'paperid': 0, 'sampleid': 0, 'samplename': 'x=0,y=0'},
{'composition': 'Ba24Ga4Ag0Ge96', 'paperid': 0, 'sampleid': 1, 'samplename': 'x=4,y=0'},
...
```

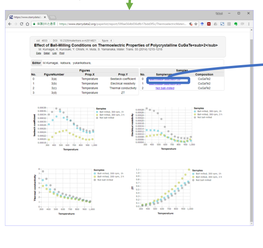

clathrate.json
149 papers / 623 figures / 890 samples
73281 data points
File size: 10 MB / Download time: ~3 min

11. Reading data in python

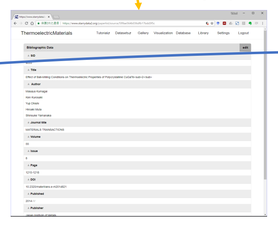


4. Mylist (custom lists of papers generated from DOI list)

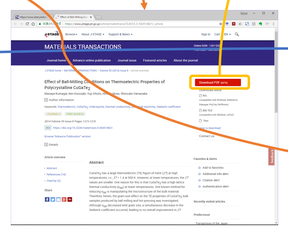
5. Report of all data



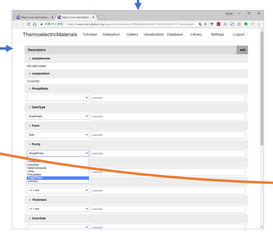
6. Details of the paper



7. Link to online PDF



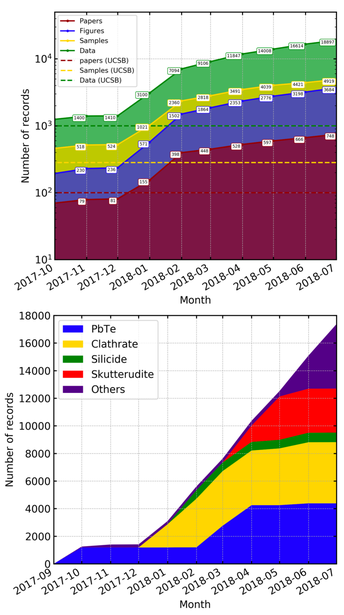
8. Sample information



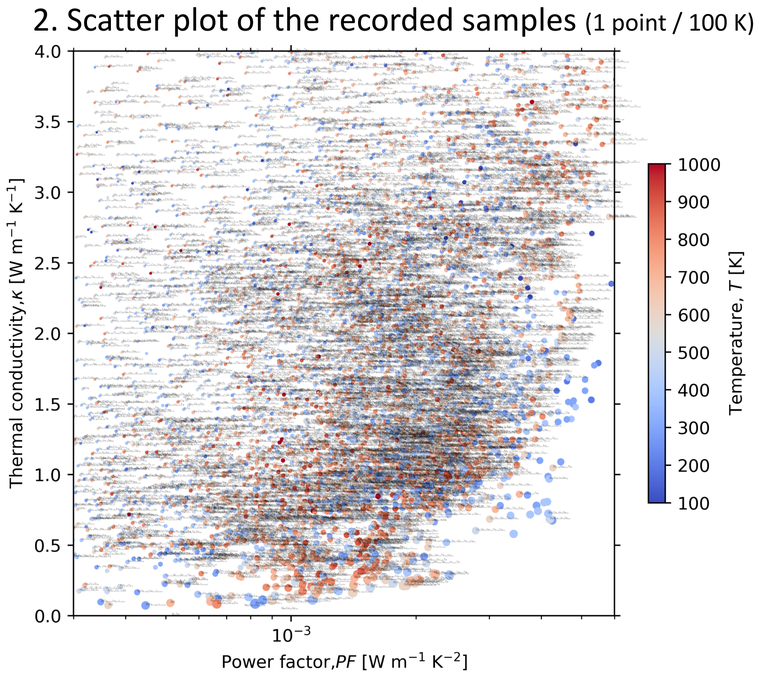
Example data: Thermoelectric properties

Thermoelectric materials are the materials that interconvert temperature difference and electricity. The efficiency increases with increasing $ZT=S^2\sigma T/\kappa$ (or increasing power factor $PF=S^2\sigma$), where S is Seebeck coefficient [V/K], σ is electrical conductivity [$\Omega^{-1}m^{-1}$], and κ is thermal conductivity [$Wm^{-1}K^{-1}$] (T is temperature [K]). Even though all these thermoelectric properties have strong material dependence and strong sample difference, these properties were not directly comparable for researchers, because each data was stored in different plot images in different papers.

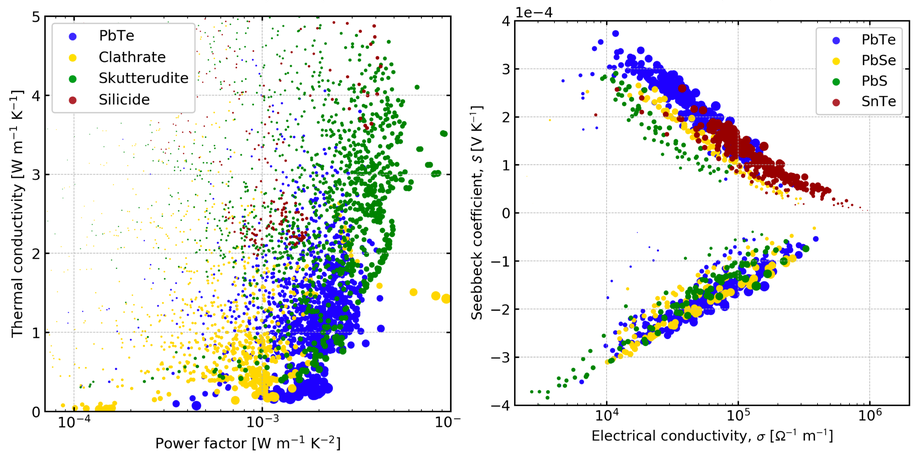
1. Data increase rate



2. Scatter plot of the recorded samples (1 point / 100 K)



3. Material family dependence of thermoelectric properties



4. Prediction by machine learning