Case study on FAIRification and certification: QsarDB

Sulev Sild
Institute of Chemistry
University of Tartu
Overview

• About QsarDB
• FAIRification
• CoreTrustSeal certification
About QsarDB

- Open access data repository for (Q)SARs models and related data
- The repository is hosted at University of Tartu (since 2012)
- The main aim is to make (Q)SAR models transparent and reusable
  - Publishing and long-term archiving
  - Citable data publications for models (DOIs)
  - Cross-link model data with journal publications
  - Web services for using deposited models/data
QsarDB details

• Data policy
  • Prefer models from scientific publications
  • Accept data in QDB archive or QMRF formats
  • Open access

• QsarDB is based on modified DSpace 6.3
  • Automated validation in data submissions
  • Visualization of models and data
  • Descriptor calculation and prediction
  • Chemical search

• Data submissions are reviewed to ensure quality
Repository contents

86 biological/chemical properties
- Phys. chem.
- Environmental fate
- Ecotoxic effects
- Human health
- Toxicokinetics
- Other

Model types
- Artificial Neural Network
- Counter-propagation Neural Network
- Decision Tree
- Ensemble Models
- k-Nearest Neighbors
- Logistic Regression
- Multi-Linear regression
- Random Forest
- Support Vector Machine

227 QDB archives
75 QMRF documents
524 QSAR models
66600 Chemical structures

227 QDB archives
75 QMRF documents
524 QSAR models
66600 Chemical structures
FAIR maturity before EOSC-Nordic collab.

• First learned about the EOSC-Nordic project in 2020
• FAIR test results from F-uji tool for QsarDB (Dec 2020)

<table>
<thead>
<tr>
<th>Name</th>
<th>Datasets</th>
<th>Platform</th>
<th>F-score</th>
<th>A-score</th>
<th>I-score</th>
<th>R-score</th>
<th>FAIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>QsarDB</td>
<td>20</td>
<td></td>
<td>42.86%</td>
<td>60.00%</td>
<td>50.00%</td>
<td>20.00%</td>
<td>36.67%</td>
</tr>
</tbody>
</table>

• May 2021 started collaboration with WP4
  • Analysed F-uji test results
  • Identified some low hanging fruit changes for improving the score
FAIRification results

• F-uji tool was very helpful
  • Very detailed output
  • Good support

• Reviewed metadata in QsarDB
  • Metadata was already in good state
  • Its machine readability was a problem
  • Improved consistency
  • Improved metadata in DataCite

• Hardest part was to configure/modify DSpace to make metadata machine readable

• We still have some room for improvement!

<table>
<thead>
<tr>
<th></th>
<th>Score earned</th>
<th>Fair level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Findable:</td>
<td>7 of 7</td>
<td>advanced</td>
</tr>
<tr>
<td>Accessible:</td>
<td>3 of 3</td>
<td>advanced</td>
</tr>
<tr>
<td>Interoperable:</td>
<td>2 of 4</td>
<td>moderate</td>
</tr>
<tr>
<td>Reusable:</td>
<td>5 of 10</td>
<td>moderate</td>
</tr>
</tbody>
</table>
CoreTrustSeal data repository certification

- We started the process started in 2021
  - Prepared the self-assessment document
  - Improved documentation on the homepage
  - Documented procedures
- EOSC-Nordic provided supported via
  - Zoom meetings
  - Reviewed the self-assessment document
- Documents were submitted to CoreTrustSeal this year
  - Received feedback
  - Some minor changes are needed
  - Resubmission is planned for this autumn
Summary

- Our collaboration with EOSC-Nordic has been valuable and productive
- With some help and a small development effort we managed to significantly improve FAIR maturity level in the QsarDB repository
- CoreTrustSeal certification process was much more difficult and it’s moving on
Thank you!

Acknowledgements

• Uko Maran (QsarDB)
• Geven Piir (QsarDB)
• Robert Huber (PANGEA)
• Andreas Jaunsen (NordForsk)
• Mari Kleemola (Tampere University)
• Mari-Elisa Kuusniemi (Helsinki University)