

This presentation is licensed under CC-BY-ND 4.0

# MeDaX - a knowledge graph for biomedicine

Judith Wodke

"Beyond Silos: Next Steps in Research Data Management"  
Frühjahrstreffen der FG Datenbanken  
Jena, 11th of March, 2024



- Diverse stakeholders



# Clinical care data

- Diverse stakeholders

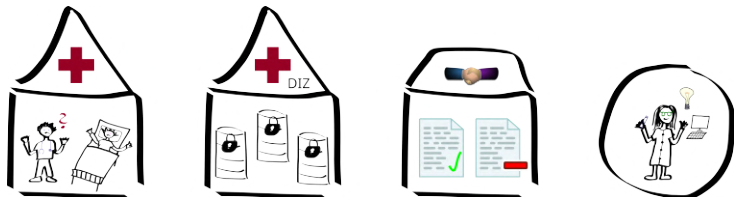


- Complex heterogeneous data



# Clinical care data

- Diverse stakeholders

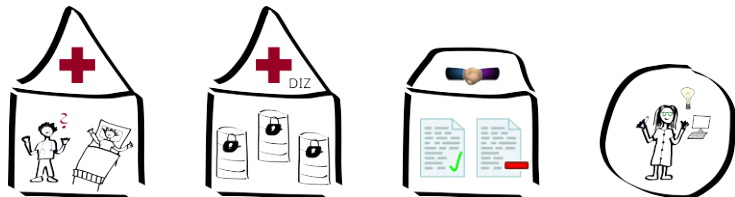


- Complex heterogeneous data
- Juridical regulations on different levels (local, national, international)



# Clinical care data

- Diverse stakeholders



- Complex heterogeneous data
- Juridical regulations on different levels (local, national, international)
- clinical data quality != experimental data quality



# Beyond Silos: Medical Informatics Initiative (MII)

- **aim:** providing medical care data for secondary use in research



# Beyond Silos: Medical Informatics Initiative (MII)

- **aim:** providing medical care data for secondary use in research
- **achievements:**



# Beyond Silos: Medical Informatics Initiative (MII)

- **aim:** providing medical care data for secondary use in research
- **achievements:**
  - federated data storage structure





# Beyond Silos: Medical Informatics Initiative (MII)

- **aim:** providing medical care data for secondary use in research
- **achievements:**
  - federated data storage structure
  - data integration centers (DICs) at all german university clinics



# Beyond Silos: Medical Informatics Initiative (MII)

- **aim:** providing medical care data for secondary use in research
- **achievements:**
  - federated data storage structure
  - data integration centers (DICs) at all german university clinics
  - MII core data set (CDS)



# Beyond Silos: Medical Informatics Initiative (MII)

- **aim:** providing medical care data for secondary use in research
- **achievements:**
  - federated data storage structure
  - data integration centers (DICs) at all german university clinics
  - MII core data set (CDS)
  - standard data exchange format



# Beyond Silos: Medical Informatics Initiative (MII)

- **aim:** providing medical care data for secondary use in research
- **achievements:**
  - federated data storage structure
  - data integration centers (DICs) at all german university clinics
  - MII core data set (CDS)
  - standard data exchange format
  - standardised shared technical infrastructure



# Beyond Silos: Medical Informatics Initiative (MII)

- **aim:** providing medical care data for secondary use in research
- **achievements:**
  - federated data storage structure
  - data integration centers (DICs) at all german university clinics
  - MII core data set (CDS)
  - standard data exchange format
  - standardised shared technical infrastructure
  - Forschungsdatenportal für Gesundheit (FDPG)

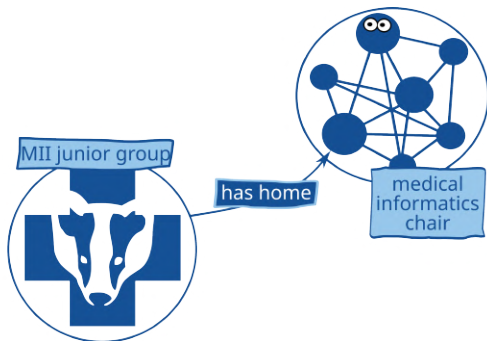


# Beyond Silos: Medical Informatics Initiative (MII)

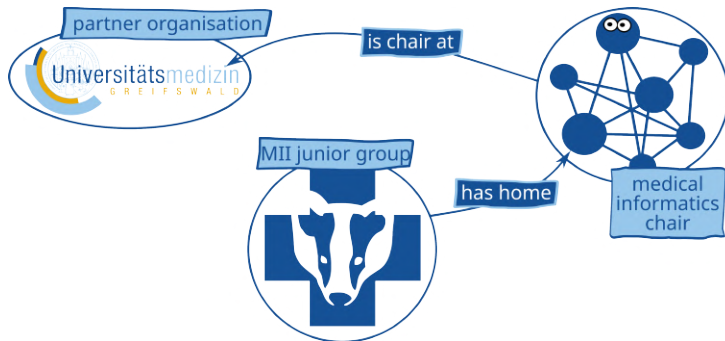
- **aim:** providing medical care data for secondary use in research
- **achievements:**
  - federated data storage structure
  - data integration centers (DICs) at all german university clinics
  - MII core data set (CDS)
  - standard data exchange format
  - standardised shared technical infrastructure
  - Forschungsdatenportal für Gesundheit (FDPG)
- **note:** work in progress, slow process, highly regulated, and many domain experts



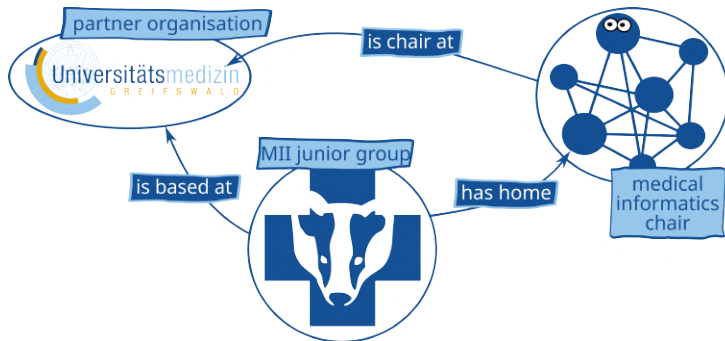


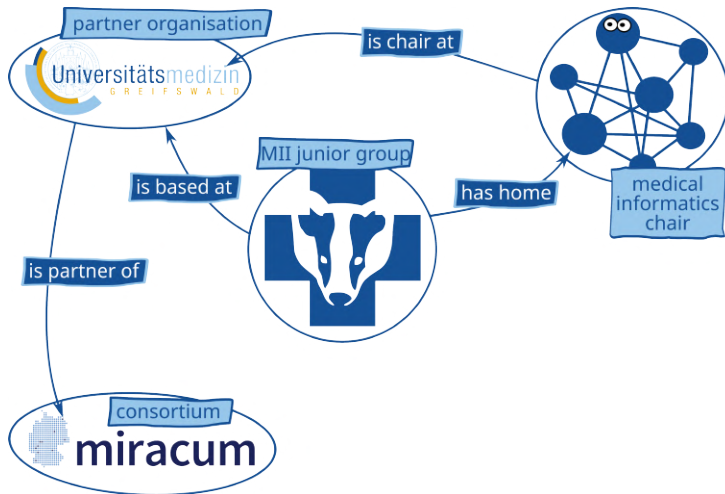


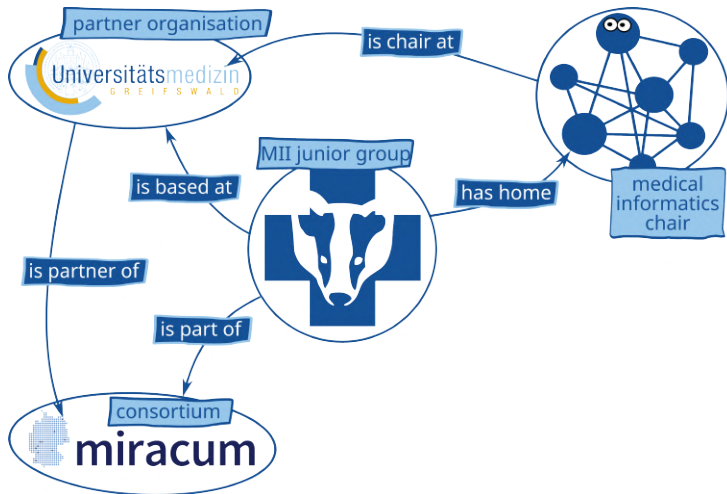


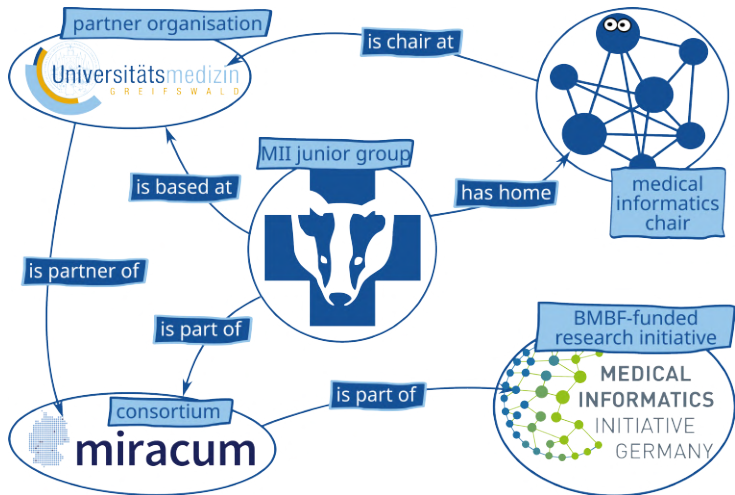


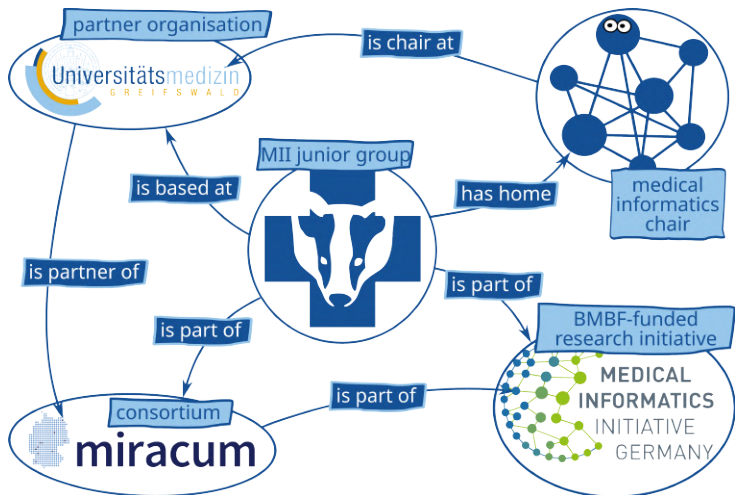
# MeDaX - bioMedical Data eXploration











# I believe in standards and reproducibility!

- **background:** from bioinformatics / theoretical biophysics (systems biology) to medical informatics:



# I believe in standards and reproducibility!

- **background:** from bioinformatics / theoretical biophysics (systems biology) to medical informatics:
  - clinics use highly diverse primary systems to collect data





# I believe in standards and reproducibility!

- **background:** from bioinformatics / theoretical biophysics (systems biology) to medical informatics:
  - clinics use highly diverse primary systems to collect data
  - consistently, data processing pipelines are highly diverse



# I believe in standards and reproducibility!

- **background:** from bioinformatics / theoretical biophysics (systems biology) to medical informatics:
  - clinics use highly diverse primary systems to collect data
  - consistently, data processing pipelines are highly diverse
  - data are formatted in FHIR according to MII CDS specifications



# I believe in standards and reproducibility!

- **background:** from bioinformatics / theoretical biophysics (systems biology) to medical informatics:
  - clinics use highly diverse primary systems to collect data
  - consistently, data processing pipelines are highly diverse
  - data are formatted in FHIR according to MII CDS specifications
  - FHIR compliance checks assure usability of data in any(!) research



# I believe in standards and reproducibility!

- **background:** from bioinformatics / theoretical biophysics (systems biology) to medical informatics:
  - clinics use highly diverse primary systems to collect data
  - consistently, data processing pipelines are highly diverse
  - data are formatted in FHIR according to MII CDS specifications
  - FHIR compliance checks assure usability of data in any(!) research
- **problem:** without standards for data collection and processing pipelines, integration of data from different sources is generally questionable



Personalised avatars by Tom Gebhardt, originals by oksmith found at [openclipart.org](https://openclipart.org)



# I believe in standards and reproducibility!

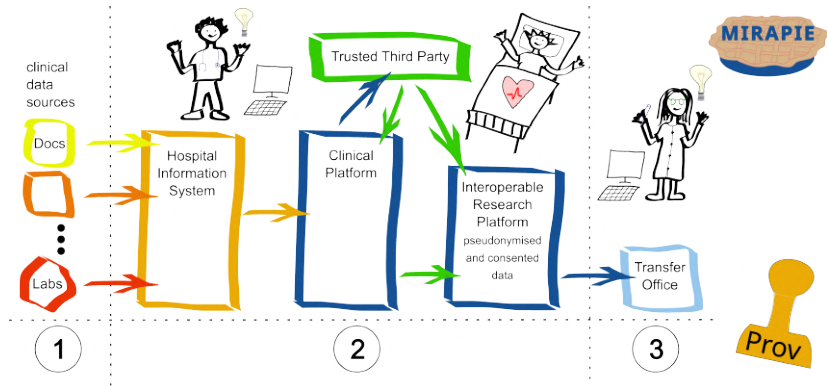
- **background:** from bioinformatics / theoretical biophysics (systems biology) to medical informatics:
  - clinics use highly diverse primary systems to collect data
  - consistently, data processing pipelines are highly diverse
  - data are formatted in FHIR according to MII CDS specifications
  - FHIR compliance checks assure usability of data in any(!) research
- **problem:** without standards for data collection and processing pipelines, integration of data from different sources is generally questionable
- **solution:** harmonise standardisation efforts and aim for international interoperability



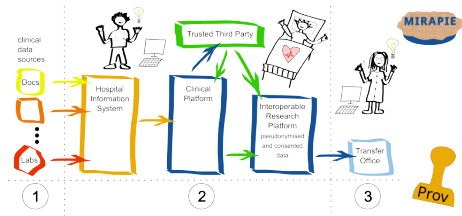
Personalised avatars by Tom Gebhardt, originals by oksmith found at [openclipart.org](https://openclipart.org)



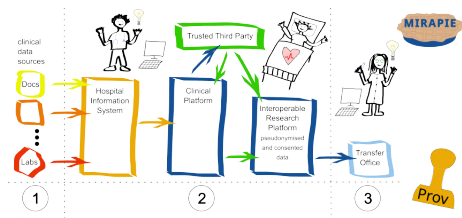
# Provenance standardisation efforts



## Minimal Requirements for Automated Provenance Information Enrichment



## Minimal Requirements for Automated Provenance Information Enrichment

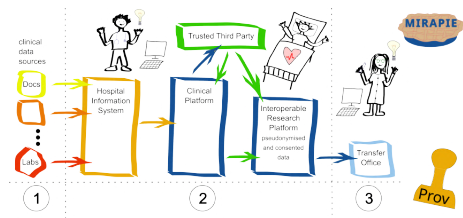


- Public repository: <https://codeberg.org/MIRAPIE/MIRAPIE>





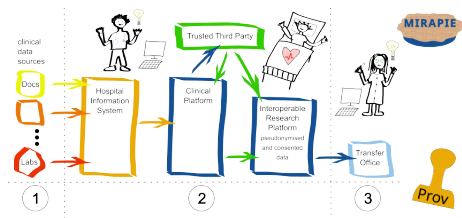
## MIInimal Requirements for Automated Provenance Information Enrichment



- Public repository: <https://codeberg.org/MIRAPIE/MIRAPIE>
- Public announcement at Provenance Week 2023 (ACM Web Conference)



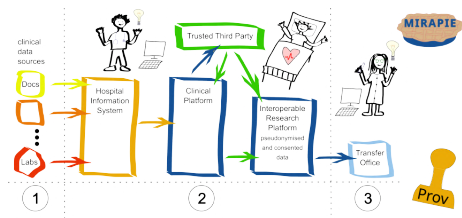
## MIInimal Requirements for Automated Provenance Information Enrichment



- Public repository: <https://codeberg.org/MIRAPIE/MIRAPIE>
- Public announcement at Provenance Week 2023 (ACM Web Conference)
- cooperate with Swiss Personalized Health Network (SPHN)



## MIInimal Requirements for Automated Provenance Information Enrichment



- Public repository: <https://codeberg.org/MIRAPIE/MIRAPIE>
- Public announcement at Provenance Week 2023 (ACM Web Conference)
- cooperate with Swiss Personalized Health Network (SPHN)
- currently onboarding french and dutch experts



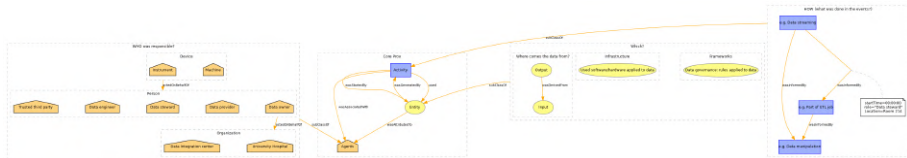
WHO does WHAT/HOW, WHERE, WHEN, WHY, and using WHICH tools with data?

- 2 more workshops in Rostock and Berlin



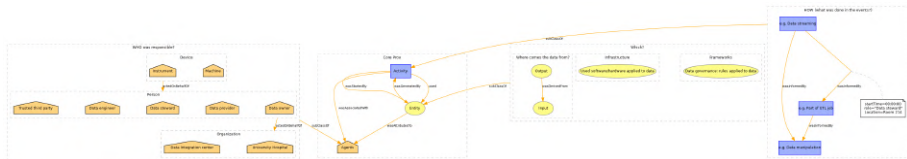
WHO does WHAT/HOW, WHERE, WHEN, WHY, and using WHICH tools with data?

- 2 more workshops in Rostock and Berlin
- a minimal data model for standardised biomedical provenance information



WHO does WHAT/HOW, WHERE, WHEN, WHY, and using WHICH tools with data?

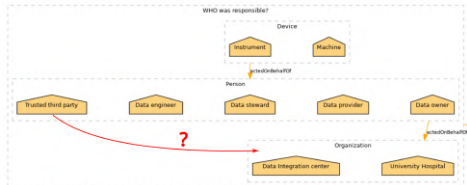
- 2 more workshops in Rostock and Berlin
- a minimal data model for standardised biomedical provenance information



- model application to use cases

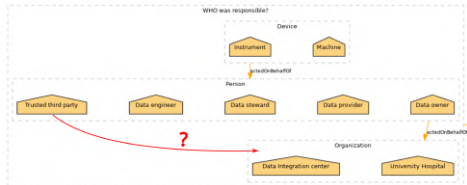
WHO does WHAT/HOW, WHERE, WHEN, WHY, and using WHICH tools with data?

- 2 more workshops in Rostock and Berlin
- a minimal data model for standardised biomedical provenance information
- model application to use cases
- remaining todos for publication:
  - finalise model



WHO does WHAT/HOW, WHERE, WHEN, WHY, and using WHICH tools with data?

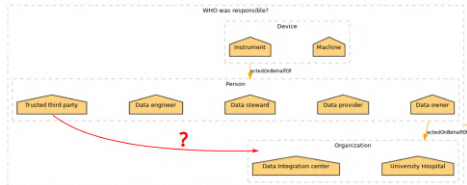
- 2 more workshops in Rostock and Berlin
- a minimal data model for standardised biomedical provenance information
- model application to use cases
- remaining todos for publication:
  - finalise model
  - apply to more theoretical and practical uses cases





WHO does WHAT/HOW, WHERE, WHEN, WHY, and using WHICH tools with data?

- 2 more workshops in Rostock and Berlin
- a minimal data model for standardised biomedical provenance information
- model application to use cases
- remaining todos for publication:
  - finalise model
  - apply to more theoretical and practical uses cases
  - acronym, figure design, and paper writing





- BC repo: <https://github.com/biocvpher/biocypher>

credits: Sebastian Lobentanzer, -TEAM, and Julio Saez-Rodriguez





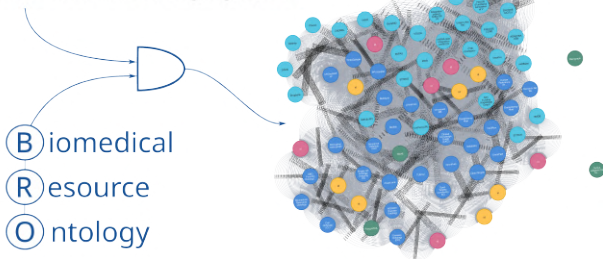
ⓑiomedical  
Ⓜesource  
Ⓞntology

- BC repo: <https://github.com/biocypher/biocypher>
- BRO repo: <https://github.com/biocypher/biomedical-resource-ontology>



# Community engagement: BioCypher + BRO

 **biocypher**  
a unifying framework for  
biomedical knowledge graphs



- BC repo: <https://github.com/biocypher/biocypher>
- BRO repo: <https://github.com/biocypher/biomedical-resource-ontology>
- meta-graph repo: <https://github.com/biocypher/meta-graph>



- background



- background
  - originally created 2010 by Jessica Tenenbaum and coworkers:  
[doi.org/10.1016/j.jbi.2010.10.003](https://doi.org/10.1016/j.jbi.2010.10.003)



- background
  - originally created 2010 by Jessica Tenenbaum and coworkers:  
[doi.org/10.1016/j.jbi.2010.10.003](https://doi.org/10.1016/j.jbi.2010.10.003)
  - not used within a decade



- background
  - originally created 2010 by Jessica Tenenbaum and coworkers:  
[doi.org/10.1016/j.jbi.2010.10.003](https://doi.org/10.1016/j.jbi.2010.10.003)
  - not used within a decade
- Jessica consented to us adopting the BRO





- background
  - originally created 2010 by Jessica Tenenbaum and coworkers:  
[doi.org/10.1016/j.jbi.2010.10.003](https://doi.org/10.1016/j.jbi.2010.10.003)
  - not used within a decade
- Jessica consented to us adopting the BRO
- moving the BRO from bioportal to the BC project (v4.0.0):  
<https://github.com/biocypher/biomedical-resource-ontology>



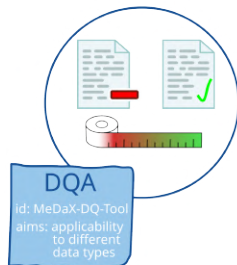
- background
  - originally created 2010 by Jessica Tenenbaum and coworkers:  
[doi.org/10.1016/j.jbi.2010.10.003](https://doi.org/10.1016/j.jbi.2010.10.003)
  - not used within a decade
- Jessica consented to us adopting the BRO
- moving the BRO from bioportal to the BC project (v4.0.0):  
<https://github.com/biocypher/biomedical-resource-ontology>
- reducing it to its core (biomedical resources) and adding the adapter class



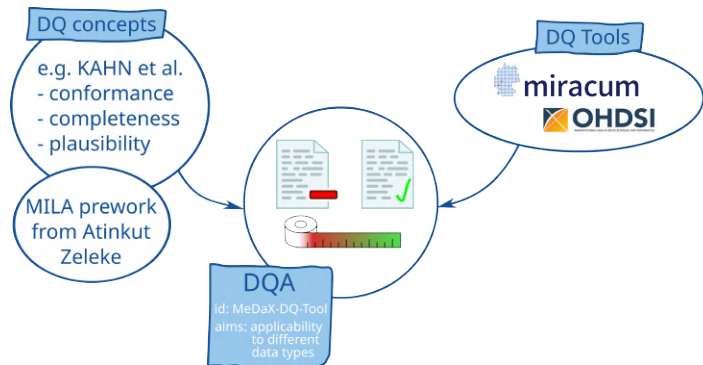
- background
  - originally created 2010 by Jessica Tenenbaum and coworkers:  
[doi.org/10.1016/j.jbi.2010.10.003](https://doi.org/10.1016/j.jbi.2010.10.003)
  - not used within a decade
- Jessica consented to us adopting the BRO
- moving the BRO from bioportal to the BC project (v4.0.0):  
<https://github.com/biocypher/biomedical-resource-ontology>
- reducing it to its core (biomedical resources) and adding the adapter class
- cleaning up, removing redundancies, refining definitions, etc.



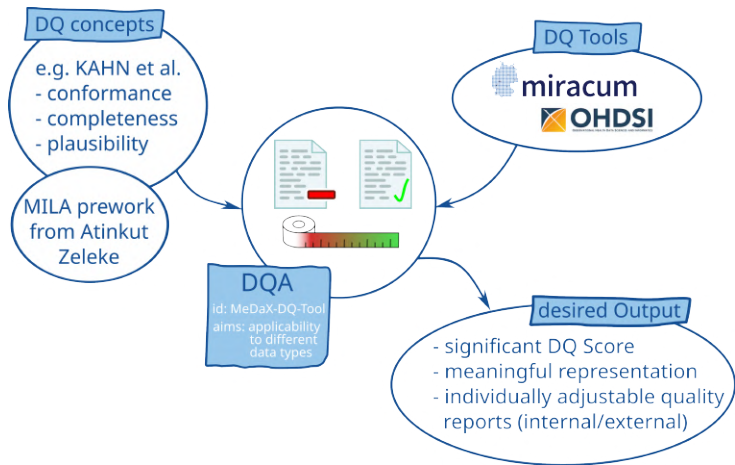
# Cooperation on Data quality (DQ) measures



# Cooperation on Data quality (DQ) measures



# Cooperation on Data quality (DQ) measures



# FAIR Impact: a FAIR assessment of the MII CDS

- FAIR Impact workshop series: baseline assessment of module *person*: exceptionally high score

|          |  |                                     |
|----------|--|-------------------------------------|
| <b>F</b> | F1-01M: Metadata is identified by a persistent identifier                          | <input checked="" type="checkbox"/> |
|          | F1-01D: Data is identified by a persistent identifier                              | <input checked="" type="checkbox"/> |
|          | F1-02M: Metadata is identified by a globally unique identifier                     | <input checked="" type="checkbox"/> |
| ⋮        |  |                                     |
| <b>A</b> | A1-01M: Metadata contains information to enable the user to get access to the data | <input checked="" type="checkbox"/> |
|          | A1-02M: Metadata can be accessed manually (i.e. with human intervention)           | <input checked="" type="checkbox"/> |
|          | A1-02D: Data can be accessed manually (i.e. with human intervention)               | <input checked="" type="checkbox"/> |
| ⋮        |  |                                     |
| <b>I</b> | I1-01M: Metadata uses knowledge representation expressed in standardised format    | <input type="checkbox"/>            |
|          | I1-01D: Data uses knowledge representation expressed in standardised format        | <input checked="" type="checkbox"/> |
|          | I1-02M: Metadata uses machine-understandable knowledge representation              | <input checked="" type="checkbox"/> |
| ⋮        |  |                                     |
| <b>R</b> | R1-01M: Plurality of accurate and relevant attributes are provided to allow reuse  | <input checked="" type="checkbox"/> |
|          | R1.1-02M: Metadata refers to a standard reuse licence                              | <input type="checkbox"/>            |
|          | R1.1-03M: Metadata refers to a machine-understandable reuse licence                | <input type="checkbox"/>            |
| ⋮        |  |                                     |

credits: Lea Michaelis, Esther Inau, Michael Muzoora, Thomas Ganslandt, Sylvia Thun, Dagmar Waltemath



# FAIR Impact: a FAIR assessment of the MII CDS

- FAIR Impact workshop series: baseline assessment of module *person*: exceptionally high score



credits: Lea Michaelis, Esther Inau, Michael Muzoora, Thomas Ganslandt, Sylvia Thun, Dagmar Waltemath





# FAIR Impact: a FAIR assessment of the MII CDS

- FAIR Impact workshop series: baseline assessment of module *person*: exceptionally high score



- next step: FAIRify the MII CDS together with TF Kerndatensatz

credits: Lea Michaelis, Esther Inau, Michael Muzoora, Thomas Ganslandt, Sylvia Thun, Dagmar Waltemath



# FAIR Impact: a FAIR assessment of the MII CDS

- FAIR Impact workshop series: baseline assessment of module *person*: exceptionally high score



- next step: FAIRify the MII CDS together with TF Kerndatensatz
- note: different levels of and purposes for tracing of data quality

credits: Lea Michaelis, Esther Inau, Michael Muzoora, Thomas Ganslandt, Sylvia Thun, Dagmar Waltemath



# The MeDaX-KG v0.2 (prototype)

Step 1: FHIR formatted research data is generically converted into a neo4j graph database: reuse CyFHIR

```
{
  "id": "1",
  "url": "http://hl7.org/fhir/us/core/StructureDefinition/us-core-patient",
  "name": [
    {
      "type": "string",
      "value": "John Doe"
    }
  ],
  "gender": "male",
  "birthdate": "1970-01-01",
  "social-security-number": "123-45-6789",
  "patient": [
    {
      "type": "string",
      "value": "John Doe"
    }
  ],
  "condition": [
    {
      "type": "string",
      "value": "Hypertension"
    }
  ],
  "medication": [
    {
      "type": "string",
      "value": "Lisinopril"
    }
  ],
  "allergy": [
    {
      "type": "string",
      "value": "Penicillin"
    }
  ],
  "immunization": [
    {
      "type": "string",
      "value": "MMR"
    }
  ],
  "test": [
    {
      "type": "string",
      "value": "Hemoglobin A1c"
    }
  ],
  "testresult": [
    {
      "type": "string",
      "value": "5.7"
    }
  ],
  "procedure": [
    {
      "type": "string",
      "value": "Chest X-ray"
    }
  ],
  "observation": [
    {
      "type": "string",
      "value": "Blood Pressure"
    }
  ],
  "observationvalue": [
    {
      "type": "string",
      "value": "120/80"
    }
  ],
  "location": [
    {
      "type": "string",
      "value": "Johns Hopkins Hospital"
    }
  ],
  "locationaddress": [
    {
      "type": "string",
      "value": "725 North Wolfe Street"
    }
  ],
  "locationcity": [
    {
      "type": "string",
      "value": "Baltimore"
    }
  ],
  "locationstate": [
    {
      "type": "string",
      "value": "MD"
    }
  ],
  "locationzip": [
    {
      "type": "string",
      "value": "21285"
    }
  ],
  "locationcountry": [
    {
      "type": "string",
      "value": "USA"
    }
  ],
  "locationpostalcode": [
    {
      "type": "string",
      "value": "21285"
    }
  ],
  "locationcountrycode": [
    {
      "type": "string",
      "value": "USA"
    }
  ]
}
```

credits: CyFHIR repository: <https://github.com/Optum/CyFHIR/>



# The MeDaX-KG v0.2 (prototype)

Step 1: FHIR formatted research data is generically converted into a neo4j graph database: reuse CyFHIR

```

{
  "id": "1",
  "type": "Patient",
  "name": "John Doe",
  "gender": "Male",
  "birthDate": "1970-01-01",
  "socialSecurityNumber": "123-45-6789",
  "maritalStatus": "Married",
  "address": "123 Main St, Anytown, CA 94024",
  "phone": "415-555-1234",
  "email": "john.doe@example.com",
  "insurance": {
    "type": "Private",
    "name": "Blue Cross",
    "id": "BC123456789"
  },
  "allergy": {
    "type": "Allergy",
    "name": "Penicillin",
    "severity": "Severe"
  },
  "condition": {
    "type": "Condition",
    "name": "Hypertension",
    "onsetDate": "2015-06-01",
    "status": "Active"
  },
  "medication": {
    "type": "Medication",
    "name": "Lisinopril",
    "dosage": "10mg",
    "start": "2015-06-01",
    "stop": null
  },
  "device": {
    "type": "Device",
    "name": "Blood Pressure Monitor",
    "model": "BP-1000",
    "serial": "123456789"
  }
}

```



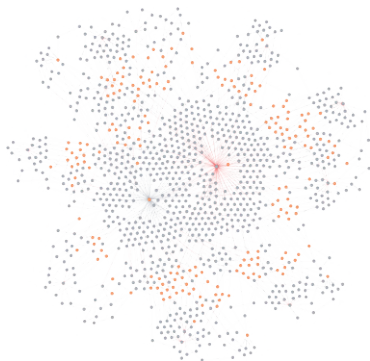
FHIR example: patient    CyFHIR data model: patient

credits: CyFHIR repository: <https://github.com/Optum/CyFHIR/>



# The MeDaX-KG v0.2 (prototype)

Step 2: Optimisation of graph granularity



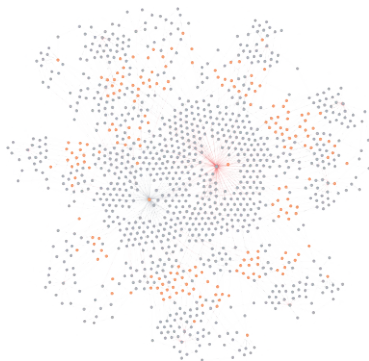
generic graph  $\rightarrow$  Neo4j

credits: Ilya Mazein

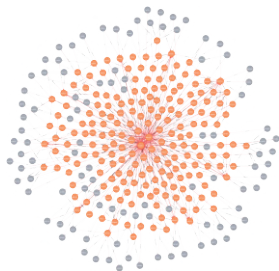


# The MeDaX-KG v0.2 (prototype)

## Step 2: Optimisation of graph granularity



generic graph → Neo4j



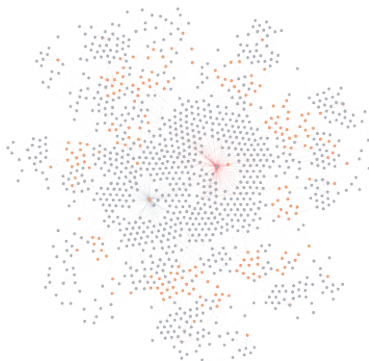
post-processed graph

credits: Ilya Mazein

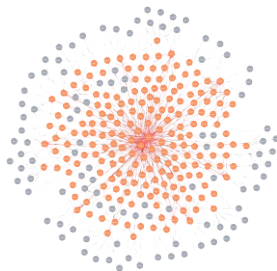


# The MeDaX-KG v0.2 (prototype)

## Step 2: Optimisation of graph granularity



generic graph → Neo4j



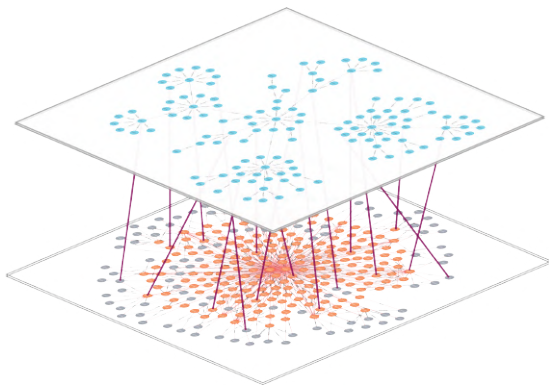
post-processed graph

credits: Ilya Mazein



# The MeDaX-KG v0.2 (prototype)

Step 3: Integration of BioLink data model: BioCypher



credits: Ilya Mazein and Tom Gebhardt





- MeDaX-KG + common data model (BioLink + BRO)

credits: Sarah Braun, Tom Gebhardt, Ilya Mazein, Lea Michaelis, Ron Henkel



# The BC-MeDaX-KG prototype - WIP

- MeDaX-KG + common data model (BioLink + BRO)
- granularity optimisation (germany-specific test data would be nice!)

credits: Sarah Braun, Tom Gebhardt, Ilya Mazein, Lea Michaelis, Ron Henkel



# The BC-MeDaX-KG prototype - WIP

- MeDaX-KG + common data model (BioLink + BRO)
- granularity optimisation (germany-specific test data would be nice!)
- automatic inclusion of new nodes and relations to BC input yaml file based on source data

credits: Sarah Braun, Tom Gebhardt, Ilya Mazein, Lea Michaelis, Ron Henkel



# The BC-MeDaX-KG prototype - WIP

- MeDaX-KG + common data model (BioLink + BRO)
- granularity optimisation (germany-specific test data would be nice!)
- automatic inclusion of new nodes and relations to BC input yaml file based on source data
- cleaning up repo, code, documentation

credits: Sarah Braun, Tom Gebhardt, Ilya Mazein, Lea Michaelis, Ron Henkel



- **in biomedicine we are working with the data of people**

credits: Benjamin Winter



- **in biomedicine we are working with the data of people**
- assuring privacy of the data owners has highest priority

credits: Benjamin Winter



- **in biomedicine we are working with the data of people**
- assuring privacy of the data owners has highest priority
- in Germany we have a federated storage structure

credits: Benjamin Winter



- **in biomedicine we are working with the data of people**
- assuring privacy of the data owners has highest priority
- in Germany we have a federated storage structure
- accordingly, we provide a dockerized MeDaX-KG pipeline that is applied locally at the DICs

credits: Benjamin Winter





- **in biomedicine we are working with the data of people**
- assuring privacy of the data owners has highest priority
- in Germany we have a federated storage structure
- accordingly, we provide a dockerized MeDaX-KG pipeline that is applied locally at the DICs
- user access control is sovereignty of the DICs

credits: Benjamin Winter



- publish our results



- publish our results
- test our MeDaX pipeline in UMGreifswald productive DIC environment



- publish our results
- test our MeDaX pipeline in UMGreifswald productive DIC environment
- implement a clinic-internal information portal



- publish our results
- test our MeDaX pipeline in UMGreifswald productive DIC environment
- implement a clinic-internal information portal
- obtain more third party funding



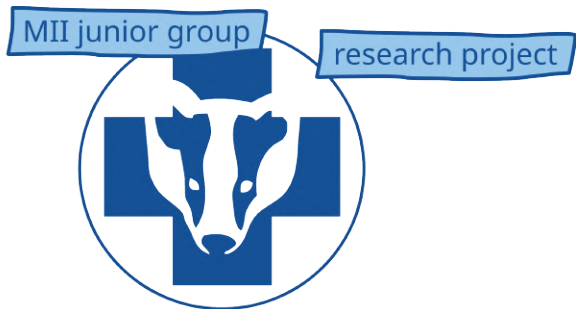
- publish our results
- test our MeDaX pipeline in UMGreifswald productive DIC environment
- implement a clinic-internal information portal
- obtain more third party funding
- integrate further data sources into our MeDaX-KG (new law in M-V upcoming: complete HIS data is potential input)



# What is MeDaX?

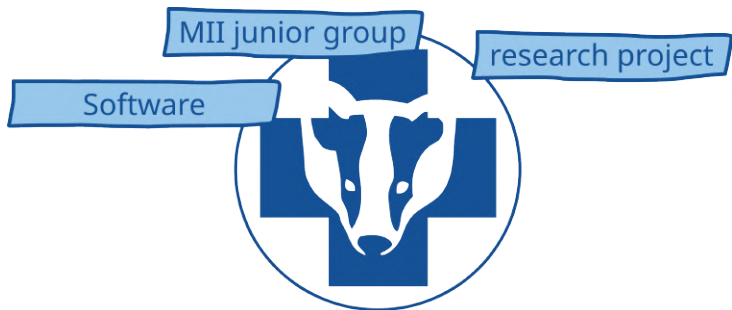


# What is MeDaX?

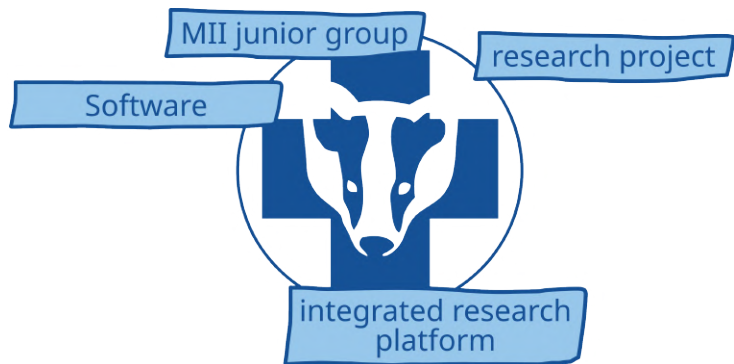




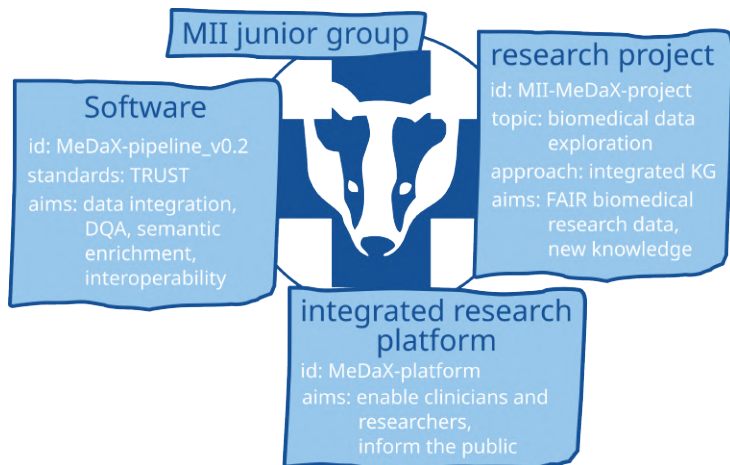
# What is MeDaX?



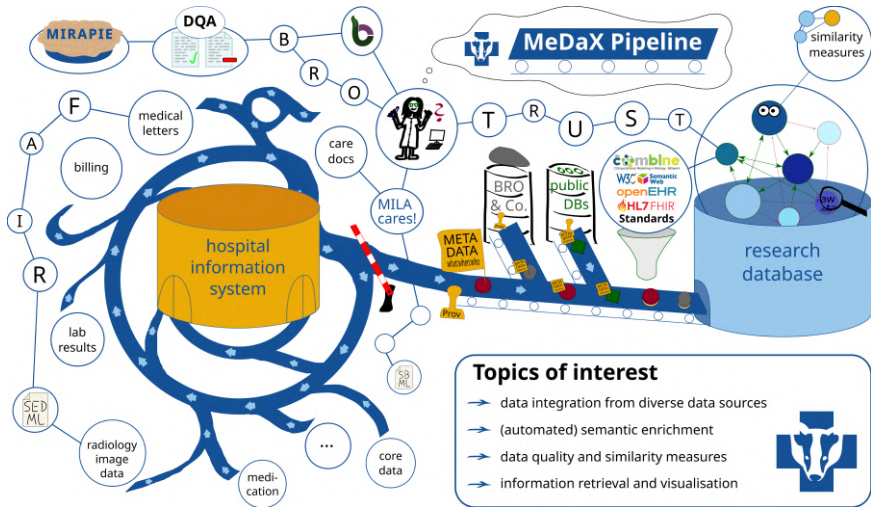
# What is MeDaX?



# What is MeDaX?

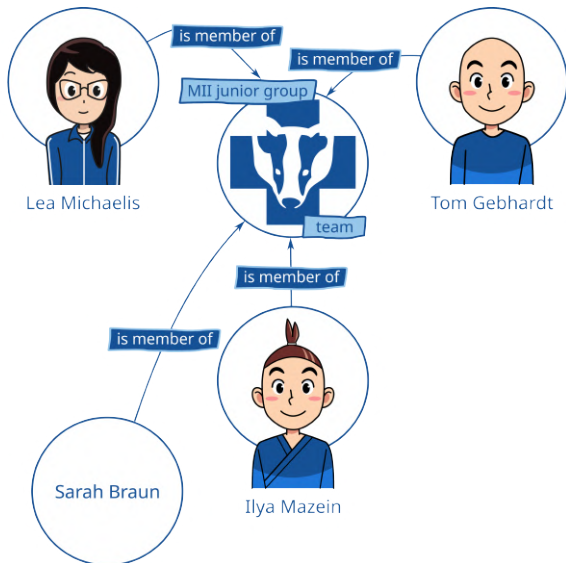


# MeDaX-Wimmelbild

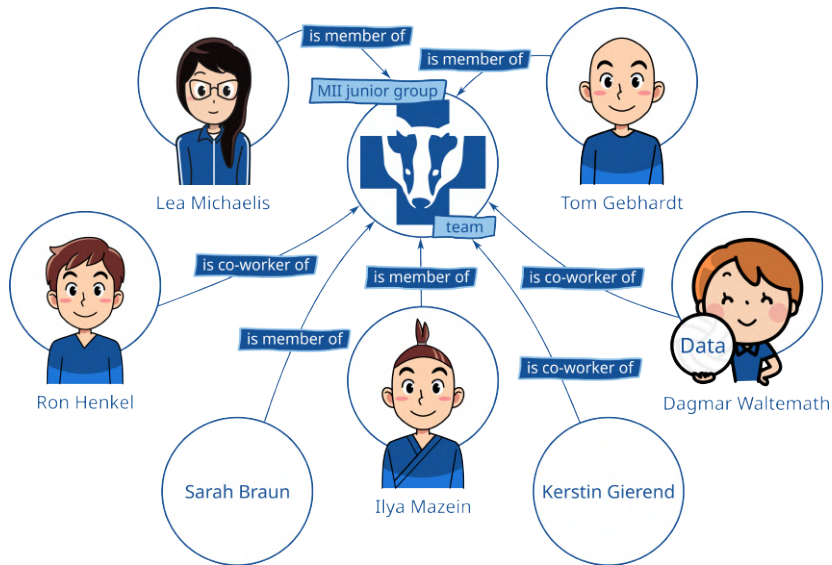




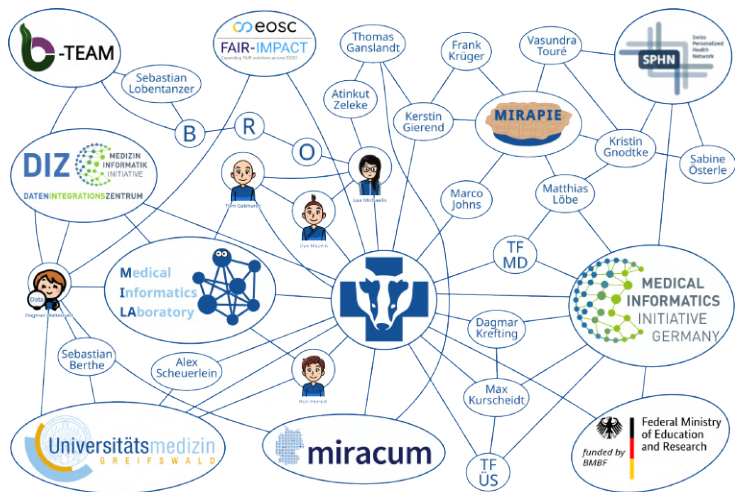
# The MeDaX-Team



# The MeDaX-Team



# Acknowledgements



Thanks for your attention!





# References and Repositories

- BioCypher paper: Lobentanzer et al., Nat. Biotech. 41, 1056–1059 (2023), doi: 10.1038/s41587-023-01848-y
- BioCypher repository: <https://github.com/biocypher/biocypher>
- BRO paper: Tenenbaum et al., J Biomed Inform 2011, 44(1):137-45, doi: 10.1016/j.jbi.2010.10.003
- BRO repository: <https://github.com/biocypher/biomedical-resources-ontology>
- CyFHIR repository: <https://github.com/Optum/CyFHIR/>
- DQA concept: Kahn et al., EGEMS (Wash DC) 2016, 4(1):1244, doi: 10.13063/2327-9214.1244 doi: 10.13063/2327-9214.1244
- Meta-Graph repository: <https://github.com/biocypher/meta-graph>
- MILA homepage: <https://www.medizin.uni-greifswald.de/medizininformatik/>
- MIRAPIE paper: Gierend et al., WWW '23 Companion: Companion Proceedings of the ACM Web Conference 2023, doi: 10.1145/3543873.3587562
- MIRAPIE repository: <https://codeberg.org/MIRAPIE/MIRAPIE>

