The e-Lab Metadata Registry: Supporting Harmonisation Using DDI, OWL and Reporting Guidelines

Dr. Chris Munro
chris.munro@manchester.ac.uk
6\textsuperscript{th} December 2016
www.herc.ac.uk & www.farrinstitute.org
HIRC's

- In August 2012, ten UK funding agencies awarded four Centres of Excellence in e-health informatics research.
- The four HIRC's aim to optimize the use of health records in research and address the UK’s capacity building requirements to support a sustainable health informatics research base.
- Manchester University hosts the Northern HIRC called the Health e-Research Centre (HeRC).
- HeRC involves collaborations at the Universities of Lancaster, Liverpool and York, and NHS partners across North England.
Farr Institute

• In 2013, the Farr Institute was created to support the HIRCs’ collective work.
  – Farr Institute @ CIPHER
  – Farr Institute @ HeRC
  – Farr Institute @ Scotland
  – Farr Institute @ UCL Partners

• Together, they bring a total of 21 academic institutions and two MRC units.
HeRC E-Lab

Aim:
• ‘Provide the missing link between data sources and research outputs by capturing the processes of research’

• Part of the Farr Institute Digital e-Infrastructure, across a number of projects

STELAR – Study Team for Early Life Asthma Research:

• UK Cohort of five follow-up studies in asthma
• It is likely that asthma is not a single disease, but a collection of several diseases
• Two recent reports have proposed that ‘asthma syndrome’ should be divided into distinct entities assigned as ‘asthma endotypes’
• Need for statistical power drives pooling of data across multiple cohorts
• Asthma-e-Lab platform will bring communities together to share data and findings
• Includes *clinical results* data and *questionnaire* follow-up data
HeRC e-Lab Platform
Levels of interaction

1. Document
   - Dropbox style functionality
2. Dataset
   - Cataloguing
3. Variable
   - Integration
   - Search
   - Reports
4. Computation

Increasing capability
Increasing semantics
Challenges

- **Phenotypic data**: need to specify relationship of columns so users understand context when searching and interpreting data.
- Data **not originally collected together** so need standard variables to pool data.
- Want curated/reusable assets at **variable level** rather than whole datasets/documents.
- Want to reuse existing terminology in describing data – e.g. SNOMED-CT, MeSH.
- Statistical packages have limitations e.g. representing the relationship between columns, metadata available.
Requirements

Core e-lab:
• document and user management
• collaboration, e.g. spaces, discussions, notifications
• authentication, authorization

Would like to use DDI to describe studies/questionnaires:
• study units, sweeps, data collection, datasets
• question items, response domains
• control constructs

To this we add:
• minimum Information Checklists (represent reporting requirements)
• support for terminologies e.g. SNOMED medical terminology

And need:
• registry to manage reporting guidelines
• software to create/edit reporting guidelines
Minimum Information Checklists*

‘Specify the minimal set of data to provide when reporting data’

For reporting a MeanWhealDiameter:

<table>
<thead>
<tr>
<th>Data Requirement</th>
<th>Requirement Level</th>
<th>Cardinality Restriction</th>
<th>Vocabulary Restriction</th>
</tr>
</thead>
<tbody>
<tr>
<td>MeanWhealDiameter</td>
<td>Primary</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>SubjectID</td>
<td>Compulsory</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Wave</td>
<td>Optional</td>
<td>0...1</td>
<td>Waves</td>
</tr>
<tr>
<td>Study</td>
<td>Compulsory</td>
<td>1</td>
<td>Studies</td>
</tr>
<tr>
<td>Allergen</td>
<td>Compulsory</td>
<td>1</td>
<td>Allergens</td>
</tr>
<tr>
<td>Description</td>
<td>Optional</td>
<td>0...1</td>
<td></td>
</tr>
</tbody>
</table>

# Retrospective Harmonization

**Property=**WhealDiameter  
**Datatype=**Integer  
**Unit=**mm  
**Wave=**1

<table>
<thead>
<tr>
<th>SubjectId</th>
<th>MWD_1</th>
<th>MWD_2</th>
<th>Spt_Sol_1</th>
<th>History_Asthma</th>
<th>Spt_Sol_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>10112</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>12034</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>12342</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>12367</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>2</td>
</tr>
<tr>
<td>16453</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>-1</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SubjectId</th>
<th>MWD_1</th>
<th>MWD_2</th>
<th>Spt_Sol_1</th>
<th>History_Asthma</th>
<th>Spt_Sol_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>1</td>
<td>2</td>
<td>&quot;celery_sol&quot;</td>
<td>&quot;Yes&quot;</td>
<td>&quot;peach&quot;</td>
</tr>
<tr>
<td>2001</td>
<td>3</td>
<td>1</td>
<td>&quot;celery_sol&quot;</td>
<td>&quot;No&quot;</td>
<td>&quot;peach&quot;</td>
</tr>
<tr>
<td>2002</td>
<td>4</td>
<td>4</td>
<td>&quot;celery_sol&quot;</td>
<td>&quot;Missing&quot;</td>
<td>&quot;peach&quot;</td>
</tr>
</tbody>
</table>

**Study A**

- SubjectId 10112: MWD_1 = 2, MWD_2 = 2, Spt_Sol_1 = 1, History_Asthma = 0, Spt_Sol_2 = 2
- SubjectId 12034: MWD_1 = 1, MWD_2 = 1, Spt_Sol_1 = 1, History_Asthma = 0, Spt_Sol_2 = 2
- SubjectId 12342: MWD_1 = 2, MWD_2 = 2, Spt_Sol_1 = 1, History_Asthma = 1, Spt_Sol_2 = 2
- SubjectId 12367: MWD_1 = 2, MWD_2 = 2, Spt_Sol_1 = 1, History_Asthma = -1, Spt_Sol_2 = 2
- SubjectId 16453: MWD_1 = 3, MWD_2 = 3, Spt_Sol_1 = 1, History_Asthma = -1, Spt_Sol_2 = 2

**Study B**

- SubjectId 2000: MWD_1 = 1, MWD_2 = 2, Spt_Sol_1 = "celery_sol", History_Asthma = "Yes", Spt_Sol_2 = "peach"
- SubjectId 2001: MWD_1 = 3, MWD_2 = 1, Spt_Sol_1 = "celery_sol", History_Asthma = "No", Spt_Sol_2 = "peach"
- SubjectId 2002: MWD_1 = 4, MWD_2 = 4, Spt_Sol_1 = "celery_sol", History_Asthma = "Missing", Spt_Sol_2 = "peach"

**Wave**  
- 1: 10/12/2013  
- 2: 10/12/2014

**Variable Report**  
- Study A: 1="Celery", 2="Egg", -1="Missing"  
- Study B: 1="Celery", 2="Egg", -1="Missing"

**MeanWhealDiameter**
### Variable Reports and Models

**Variable Report Model** (or reporting guidelines)

<table>
<thead>
<tr>
<th>Data Requirement</th>
<th>Requirement Level</th>
<th>Cardinality Restriction</th>
<th>Vocabulary Restriction</th>
</tr>
</thead>
<tbody>
<tr>
<td>MeanWhealDiameter</td>
<td>Primary</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>SubjectID</td>
<td>Compulsory</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Wave</td>
<td>Optional</td>
<td>0...1</td>
<td>Waves</td>
</tr>
<tr>
<td>Study</td>
<td>Compulsory</td>
<td>1</td>
<td>Studies</td>
</tr>
<tr>
<td>Allergen</td>
<td>Compulsory</td>
<td>1</td>
<td>Allergens</td>
</tr>
<tr>
<td>Description</td>
<td>Optional</td>
<td>0...1</td>
<td></td>
</tr>
</tbody>
</table>

**Variable Report** (reporting against the requirements)

<table>
<thead>
<tr>
<th>Data Requirement</th>
<th>Data Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MeanWhealDiameter</td>
<td>MWD_Var</td>
</tr>
<tr>
<td>SubjectID</td>
<td>SubjectID</td>
</tr>
<tr>
<td>Wave</td>
<td>“STELAR_WAVE_1”</td>
</tr>
<tr>
<td>Study</td>
<td>“STELAR”</td>
</tr>
<tr>
<td>Allergen</td>
<td>“Celery”</td>
</tr>
<tr>
<td>Description</td>
<td>“Mean wheal diameter – protocol ...”</td>
</tr>
</tbody>
</table>
### Variable Bank

#### STELAR

**Cohorts**

- **SEATON**
  - **IsleOfWight**

**Datasets**

- IOWAgeinWeeks
- IWRecruitment
- IYear1
- IYear4SPT
- IYear18SPT
- IWEarlySPT
- IYear2
- IYear4
- IYear10
- IYear10SPT
- SCORAD
- IWBirthFactors
- Spirometry18
- Spirometry10
- IW_SC_Harmonisation
- MethacholineChallenge

- **MAAS**
- **ALSPAC**
- **ASHFORD**
- **New Cohort**

#### Displaying page 1 of 1 at 50 variables per page. 0 of 11 variables are selected

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Template</th>
<th>Label</th>
<th>FollowUp</th>
<th>ID</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOWAgeinWeeks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IWRecruitment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IYear1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IYear4SPT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IYear18SPT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IWEarlySPT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IYear2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IYear4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IYear10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IYear10SPT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCORAD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IWBirthFactors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spirometry18</td>
<td><strong>PostBronchodilator</strong></td>
<td>Postbronchodilator FEV1</td>
<td>FollowUp.IWAge18Y</td>
<td>spirometry18.FEV1_Post_18</td>
<td></td>
</tr>
<tr>
<td>Spirometry10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IW_SC_Harmonisation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MethacholineChallenge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAAS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALSPAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASHFORD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New Cohort</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Variable Browse

#### Cohorts
- **ASHFORD**
  - QuestionnaireYear3
  - QuestionnaireYear8
  - FieldWorkYear5
  - Pregnancy2
  - FieldWorkYear14
  - Yearonevisit
  - pregnancy1
  - QuestionnaireYear5
  - FieldWorkYear8
  - QuestionnaireYear2
  - Exposuredetails
  - Birthdetails
  - Spirometry
  - QuestionnaireYear14
  - QuestionnaireYear1
  - SkinPrickTesting1
  - SES

#### Parameters:
- **Cohort:** ASHFORD
- **Template:** Anthropometry
- **Label:** Head circumference
- **FollowUp:** FollowUp.ADBirth
- **ID:** birthdetails.HEADCIRC
- **Tags:** head circumference

#### Related Variables:
- Birthdetails.BIRTHWT, Birthdetails.LENGTH

#### Table:
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIW</td>
<td>Refers to values in data</td>
</tr>
<tr>
<td>AnthropometryType</td>
<td>AnthropometricMeasure.HeadCircumference</td>
</tr>
<tr>
<td>ClinicalType</td>
<td>ClinicalMeasurement.AnthropometricMeasure</td>
</tr>
<tr>
<td>DataAnthropometricMeasure</td>
<td>Refers to values in data</td>
</tr>
<tr>
<td>DataSource</td>
<td>DataSource.ClinicalMeasurement</td>
</tr>
<tr>
<td>DateID</td>
<td>Refers to values in data</td>
</tr>
<tr>
<td>FollowUp</td>
<td>FollowUp.ADBirth</td>
</tr>
<tr>
<td>MethodOfFollowUp</td>
<td>FollowUpMethod.ClinicVisit</td>
</tr>
<tr>
<td>Subject</td>
<td>Person.StudySubject</td>
</tr>
<tr>
<td>SubjectNo</td>
<td>Refers to values in data</td>
</tr>
</tbody>
</table>
Implications for Discovery and Reuse

- Lowering the barrier to reuse of data
  - automated data extracts once data uploaded
  - reduce manual work combining data from multiple cohorts

- Variable Reports
  - a new ‘unit of work’ / reusable standards based asset
  - searched/exported instead of entire datasets

- Incorporate additional context to variables described in DDI for retrospective harmonization

- Allow reuse of existing terminologies e.g. SNOMED within data harmonization process
Conclusion

• Variable Reports/ Models are reusable assets to specify/ incorporate data context to ensure correct interpretation
• Links together OWL, DDI, Minimum Information Models
• Metadata registry manages definitions
• Governance + software + process
• Challenges in discovering relationships between assets from multiple research data/publication portals – ‘Federated Discovery’ project e-Lab/CLOSER (CM, PC, JJ)
• Recently been evaluating FHIR – HL7 Healthcare Interoperability standard/tools