FROM RARE DATA TO FAIR DATA
WITH THE RARE DISEASE COMMUNITY

MARCO ROOS

Acknowledging rare disease patients & patient representatives, participants and teachers of all rare disease ‘Bring Your Own Data’ workshops, the members of the rare disease cross-project FAIR data task force, Annika Jacobsen, Claudio Carta, David van Enckevort, Mark Thompson, Rajaram Kaliyaperumal, Kees Burger, Luiz Bonino, Mark Wilkinson, Frederieke Ehrhart, Rachel Thompson, Lieze Thielemans, Claire Shovlin, Leo Schultze Kool, Marina Mordento, Luca Sangiorgi, Domenica Taruscio, Gary Saunders, Ana Rath, Marc Hanauer, RD-Connect and its members, the RDs GO FAIR seed group, RD-Connect, ELIXIR-NL/EXCELERATE, BBMRI-ERIC/ADOPT/NL, VWData, ODEX4All, FAIR-diCT
Rare disease research

**RARE DISEASES**

1:17 people

30 million in Europe

7% OF THE POPULATION ARE AFFECTED BY RARE DISEASES

THE EU CLASSES A DISEASE AS ‘RARE’ WHEN LESS THAN 1 IN 2000 SUFFER

OVER 7000 DISEASES BIOSAMPLES, DISEASE & PATIENT INFO, OMICS, GENOTYPE-PHENOTYPE
Patients demand *respectful* use of data

- Use our data
- Make our data as usable as possible
- Inform us
- Involve us
Example: drug prioritization for Huntington’s Disease

Eleni Mina, Kristina Hettne, Eelke van der Horst, Katerina Nosikova, Elizaveta Besedina, Katy Wolstencroft, Peter-Bram ‘t Hoen, Marco Roos

Select genes by phenotype matching in Monarch

Select drug compounds in OpenPHACTS

Filter on feasibility for treating HD

Prioritized drug compounds

Unpublished data
Pain distribution preparing for multi-source analysis

Experiment/survey/...

Data publication

Integrative analysis

Application

No pain ← pain → Max. pain

1 month

6 months
Pain distribution preparing for multi-source analysis

1 month

N x 6 months
FAIR Pain distribution preparing for multi-source analysis

Experiment/survey/...
Data publication
Integrative analysis
Application

6 months
1 month
FAIR Pain distribution preparing for multi-source analysis

- Experiment / survey / ...
- Data publication
- Integrative analysis
- Application

6 months

N x 1 month

5 months more research
start 5 months earlier
Siloes
Silos, silos, silos
Silos, silos, silos
Absolute need to enable analysis across resources
Usability of data for rare disease care and research must be brought to higher, global levels.
Challenges for analysis across rare disease resources

- **Findability**
  *First challenge*
  Orphanet and ‘Finder’ services on RD-Connect platform, bioschemas

- **Accessibility**
  *Unavoidable challenge*
  Requires ‘ELSI’ + Technology experts to automate consent checking

- **Interoperability**
  *Critical challenge at record level*
  Interdisciplinary collaboration, Ontologies & linked data, mapping to domain standards

- **Reusability**
  *Highest challenge*
  Requires F+A+I, collaboration with rare disease data quality guideline developers
FAIR data stewardship services: tools and people
(DTL, UPM, ISS, LUMC, ELIXIR-NL/UK/SL, BBMRI, RD-Connect, FAIR-dICT, ...)

Plan

Create

Publish

Find

Courses

Bring Your Own Data Workshop

FAIR Hackathons

Data linkage plan as a service

FAIR expert task forces

Project planning with stakeholders
Infrastructure 4: fair data services by tools and people
(DTL, UPM, ISS, LUMC, ELIXIR-NL/UK/SL, BBMRI, RD-Connect, FAIR-diICT, ...)

- Plan
- Create
- Publish
- Find

Courses
- Data linkage plan as a service
- FAIR expert task forces
- Project planning with stakeholders

Bring Your Own Data Workshop

FAIR Hackathons
Annual ‘Rome’ summer school & Bring Your Own Data workshop for rare disease data managers

- Data owners bring their data
- FAIRification by:
  - data linking experts
  - resource data experts
  - domain experts (MDs, patients & patient representatives)
- Discussion on FAIR project management
Make all my data FAIR in two days!
BYOD Hype cycle

Make all my data FAIR in two days!

End of day 1: confusion
BYOD Hype cycle

Make all my data FAIR in two days!

End of day 1: confusion

Day 2: discuss FAIRification project planning

Make all my data FAIR in two days!

End of day 1: confusion

Day 2: discuss FAIRification project planning
Define driving user question(s)

Pre-FAIRification analysis
- Analyze question, data and prioritize essential data subset

Define semantic model
- Conceptual model
- Available ontologies and data models
- Detailed instance model
- Global identifiers
  - Use of Identifiers.org

Transform data records
- Apply semantic model to data
- FAIR-compliant, machine readable knowledge graph representation

Define metadata
- Principles for F, R: “rich metadata”
- Includes usage policy/license
- Registration on RD-Connect platform

Deploy FAIR data resource
- FAIR Data Point: data and metadata (DTL)

FAIRification process

FAIR metrics

FAIR metrics
FAIR data point: metadata model driven REST API
Enabling information retrieval and analysis across RD resources

Used for rare diseases and plants!

I would like to know the number of samples of donors with an abnormality in head or neck in a specific region of Italy, in order to check if exposure to environmental factors is important.

In addition, I would like to see in which biobanks I can find the samples, the phenotypes associated with them, and information about the organisation(s) behind the biobanks or registries.

Driving user questions

Demonstrator UI

Biobank/registry info

(Biobank & Registry Finder)
Rare diseases BYOD outcomes

- First prototype ‘FAIR data point’ (BYOD hackathon)
- FAIRification procedure
- Cross-resource query demonstrator
- Cross-project rare disease data linkage plan

14 November 2018
Rare disease data linkage plan

Knowledge exchange

Your knowledge of your data

your DATA

Source data steward(s)
FAIR data steward(s)
Domain expert(s)

FAIR
Your knowledge

your DATA
Linkable

Cost: ~4-8PM
~40-80K Euro
Standardisation of rare disease services in collaboration with the RD communities

- **Osteogenesis imperfecta**
  (EXCELERATE+BBMRI ADOPT, BOND ERN): Marina Mordenti/Luca Sangiorgi (Rizzoli hospital Bologna/BOND ERN)

- **Vascular Anomalies**
  (RadboudUMC funds) Leo Schultze Kool/Karlijn Groenen (RadboudUMC), Claire Shovlin/LiezeThielemans (Imperial College London) (VASCERN), Castor EDC

- **Rare muscular dystrophies**
  (EXCELERATE+RD-Connect) LUMC Neurology department+Library+Advanced Data Management group & Rachel Thompson (Newcastle University)

- **Rett Syndrome variants**
  (ELIXIR) Rett organisations and Frederieke Ehrhart (UM)
FAIRification: Interdisciplinary learning experience

- Disease experts (MDs, patient representatives)
- “Local” rare disease data stewards
- Rare disease FAIR data stewards
- Registry software providers
- FAIR software engineers
- Cross-project rare disease FAIR data task force
LUMC FAIR data stewardship

LUMC librarians and Data Managers

LUMC ICT

LUMC FAIR data stewards

FAIR software engineers
Current challenge: scaling up
Reuse human & machine readable knowledge
Rare disease community highly engaged in establishing FAIR

FAIRification done by humans: data stewards, interdisciplinary collaboration

Findability first bottleneck, data interoperability most important bottleneck

Next challenge: scaling up
Engage!

RD-Connect community, RDs GO FAIR implementation network

Foster the adoption of FAIR data principles by the rare disease community

Ensure that rare disease patients and patient representatives are engaged.

Ensure that FAIR sharing in the rare disease community is respectful and responsible towards rare disease patients.

Join our community today!

https://rd-connect.eu/about-rd-connect/become-member/

fair-rd-info@elixir-europe.org
Thank you

FAIR data engineering team
Rajaram Kaliyaperumal, Kees Burger, Nuno Nunes, Shamanou van Leeuwen, Mark Thompson, Luiz Bonino, Mark Wilkinson, Michel Dumontier

RDs GO FAIR seed group:
Claudio Carta, Ronald Cornet, David van Enckevoort, Marc Hanauer, Ian Harrow, Victoria Hedley, Gulcin Gmus, Dipak Kalra, Veronica Popa, Ana Rath, Marco Roos, Yaffa Rubinstein, Gary Saunders, Rachel Thompson, Mark Wilkinson

Training material:
ELIXIR, FAIR-dICT, DTL, Claudio Carta, Mascha Jansen, Celia van Gelder, Erik Schultes, Albert Mons

Use case collaborators
Rachel Thompson, Libby Wood, Claudio Carta, Marco Crimi, Estrella Gomes, Marina Mordenti, Luca Sangiorgi, David van Enckevoort, Fleur Kelpin, Claire Shovlin, Lieze Thieleman, Hevas, Leo Schultze Kool, Caroline van den Bosch, Carine van Vleuten, Erik N, VSOP, Annika Jacobsen, Freddie Ehrhart, Andrea Waagmeester; Heimo Muller, Robert Reihs; Pedro Sernadella, Jose Oliveira; Marc Hanauer, Ana Rath; Roxana Merino; Matthias Brochhausen; Developers of Castor, RDRF, OSSE, MolGenis; participants Bring Your Own data workshops and patient representatives, rare disease linked data and ontology task force

ELIXIR: Chris Evelo and team (Maastricht), Carole Goble and team (Manchester), Ivo Gut and team (CNAG), Helen Parkinson and team (EBI), Olivier Poch and team (Strasbourg), Orphanet, Brane Leskosek and team (Ljubljana)

https://rd-connect.eu/about-rd-connect/become-member/

The BioSemantics group and the Human Genetics Department LUMC