Data strategies for research infrastructures
Living Document

Link to Living Document: http://tinyurl.com/kewds

Webpage

http://www.biomedbridges.eu/trainings/knowledge-exchange-workshop-data-strategies-research-infrastructures

- Slides will be uploaded here as we go
- Please send slides before you present to tom@ebi.ac.uk

Programme

09:00-09:15 Arrival and registration

09:15-09:20 Welcome and introductions (*Tom Hancocks, BioMedBridges*)

09:20-09:30 Introduction to the workshop (*Stephanie Suhr, BioMedBridges*)

09:30-09:50 Scientific & technical requirements (*Rafael Jimenez, ELIXIR*)

09:50-10:10 Financial requirements (*Steven Newhouse, EMBL-EBI - TBC*)

10:10-10:30 Formal requirements (*Stephanie Suhr*)

10:30-11:00 Data strategy examples:

--Strategies, Challenges and Prospect of Data Management at ESO (Michael F. Sterzik, ESO)

-- Emerging infrastructure (*TBC*)

11:00-11:15 Break

11:15-12:30 Breakout session 1: Understanding your RI data management requirements

12:30-13:00 Report back and discussion

13:00-14:00 Lunch

14:00-15:00 Breakout session 2: Understanding the needs of data users

15:00-15:30 Report back and discussion

15:30-15:45 Break

15:45-16:45 Flash presentations (by workshop participants)

16:45-17:30 Wrap-up discussion

Resources

- Principles of data management and sharing at European Research Infrastructures
    - doi: 10.5281/zenodo.8304 (link is external)
- REPORT: BioMedBridges workshop on e-Infrastructure support for the life sciences – Preparing for the data deluge
    - doi:10.5281/zenodo.13942(link is external)
- DRAFT e-IRG White Paper 2014
    - Best practices for the use of e-infrastructures by large-scale research infrastructures

Introduction to the workshop

Steffi Suhr, ELIXIR/BioMedBridges

Workshop is about data STRATEGY, not data management
Need to know whether you're a data producing RI, a data managing RI, or both.

What type of RI are you?
Use case - BBMRI (biobanking) - different national policies for dealing with patient data; have to ensure protection of patient/donor data.

- Data producing
- Data analysing
- Data storing

Key question: Is the provision of data a core part of your business?

- Do users arrive, use infrastructure and then leave with their data?
  - Instruct: yes, this is the typical activity of Instruct
- Do you keep data?
  - Instruct: yes, some facilities archive raw data

Factors influencing data strategy - lots of factors impact on data strategies

- Scientific (all of you!)
- Technical (Rafa)
- Financial (Steven)
- Political
- Legal/ethical/formal (Steffi)
- Social: Structural Biologists (Instruct) have strong cultural commitment to sharing (reduced) data
- Maybe add: organizational?

Scientific & technical requirements

Rafael Jimenez, ELIXIR. PPT
"If you want to go fast, go alone
If you want to go far, go together"

Common message in documents

- e-Infra look forward to help RI

Life sciences - data production is produced in many locations

- Very diverse generation and widespread

Data generation capacity is growing beyond capacities of networks to move and archives to store

- Moving tools and compute to the data?
- More varied approaches?
- Sometimes must move data to compute!
  - "Never underestimate the bandwidth of a station wagon full of tapes hurtling down the highway" [Andrew S. Tanenbaum, *Computer Networks*, 4th ed., p. 91]
  - https://what-if.xkcd.com/31/

Data analysis slide

- Take tools to data, but need compute power
- network becomes the main problem here
- 2 solutions? More compute or more network - factors involved are many
- Data units vs processing units
- Or buy a bike and take the data to the computer!

Data growth - budget shortfall in IT

- Storage cost falls slower than generation cost

The key question is flops per byte. See the classic paper
http://arxiv.org/ftp/cs/papers/0403/0403019.pdf

Q - When can we delete the data! BIG QUESTION FOR MANY!
Comment: in many countries, there are legal archiving requirements.

Generator keeps data until it is requested

- Many problems with this

Archives can't take all the data (raw or processed!)

Life science diversity

- Many different requirements for genomics, metabolomics etc

Domain specific strategies vs generic strategies

- Start with generic that work for many
- Look towards smaller more specific approaches
- Collaborations can help here

Often have data associated with an experiment that has no home - for example, a nutrition study in which you have microarray data, which can go in arrayexpress, but you also have the results of glucose tolerance tests; where do these data go?
stakeholders in data sharing -
Data producer
Data resource
Data consumer
Gergely - there may also be a fourth - the infrastructure provider

Middle tier is more complex

Financial considerations

Steven Newhouse, EMBL-EBI

A recipe for happiness - http://en.wikipedia.org/wiki/Wilkins_Micawber

- Who worries about where the money comes from to pay for equipment!
  - Not the scientist!
  - It is the managers
- European Grid Infrastructure
  - Transition from project to infrastructure
  - Funding own activities
- EBI
  - Embassy Cloud as a commercial services
  - Cost of Technical Services at EBI
- Expenditure and income are the only things you need to worry about!
  - Capital costs - building, hardware, software
  - Operating costs - People, software, support, Physical/Electronic infrastructure
    - This is more tricky in an academic setting
    - Support is a big operating cost at time
    - cloud is cheap, but helping users is costly
- Embassy Cloud (is IaaS) - commercial priced for compute next to EBI data
  - http://www.ebi.ac.uk/services
  - Has ‘state’ funding so this has been tricky to set up on a commercial basis
    - Can’t be seen as anti-competitive
    - So not a threat to AWS
- EGI Core Services
  - http://www.egi.eu/
  - Essential, useful, nice to have, not needed
  - Use community money to fund essential services
  - Project funding to develop services and add new offerings
    - Things funders might want to pay for
- EBI Technical Services
  - Understanding driving costs
  - Metrics - people, network, machines
  - Can you change behaviour of users?
    - Need incentives for research teams to move data from fast storage
      - More to tape storage?
    - EMBL Heidelberg charges research groups for storage!
      - Pushes data to long-term archives
    - EBI has only 35BT of tape storage! EBI doesn’t charge storage
  - Capital Expenditure is main funding method

this is done very quickly and is tricky!
Not great for long-term strategy
Need to move to operational expenditure
Keep a plan for one off spending if there is spare money!

Data storage cost is most in first 2 years
Next 8 years costs is roughly the same
Capitalise on falling costs
It wont last forever!

Shifting costs - centre to nodes, nodes to centre, infrastructure to users?

Costs for data storage/management should be considered part of cost for data production, in order to have a sustainable, scalable system.

Compare data users to PhD students and a free buffet…!

Sustainability is not another EC project
Look to community to provide

==========

Most of us don't know how much is costs to deliver our services
Have to think about…

Capital costs - buildings, hardware, software
Operating costs - people, software (e.g. ongoing licencing costs), support, physical and electronic infrastructure

Cloud infrastructure might be cheap to operate but your users may not be used to using it so they need support - support costs are unpredictable and distributed.

Move the computer power next to the data.

Make it very clear what your service offering is and write down the service description.

Divide services into essential, useful, nice to have, not needed. You may find there's something that you've been running for year's that no-one wants (EGI did).
Core community (e.g. members of your infrastructure) pays for essential services
Nice-to-haves should be paid for by project funding
Can change the behavior of your users through charging mechanism - EMBL HD charges researchers for storage in fast and slow tiers; this optimises use of fast storage, which is the most expensive to maintain.

If all your funding agency wants to do is give you a big CapEx budget that has to be spent in a limited time period, how do you deal with ongoing operational costs?

Once you have a budget that balances, the next question is the scaling law. Research expenditure is roughly constant. Transistors per euro, and magnetic domains per euro, are also now constant. But data acquisition instruments keep getting better.
http://blog.dshr.org/2014/05/talk-at-seagate.html

Costs of data storage obey the 80:20 rule - 80% of your costs occur in the first two years; thereafter costs drop.

Conclusions
Define a budget - researchers want everything for free but if you tell them it'll cost their priorities may change.
Sustainability is not another EC project; you have to look to the community to provide funds for what is essential to it.

Formal requirements

Stephanie Suhr, BioMedBridges

Political factors
- Funding bodies
- Large organisations
- National interests

Don't get hung up on technical or scientific factors
- What are the politicians wanting?
- What is the aim of the infrastructure from political entities POV

Legal factors
- Patient data/health records - country of origin laws
- Language - medical records, legal documentations

Formal factors
- EC - wants access to and reuse of data, prioritise dissemination of data
- H2020 - open data pilot? Opt in method for this is way forward

National funding bodies
- UK and NL
- US funding bodies looking for open-access of data with government funding

IP of data access
- Private/public partnerships
- Commercial/private entities

Public research
Funded through the RI (internal projects)

Sensitive data
ELSI
IP
software/data licenses

Q - where does rare disease community sit?
It is a rather open community, need to find patients with similar conditions

Making sensitive data available - open and transparent about who has access and accountability, security and informed consent and acknowledgement, legal frameworks

Social factors - culture of the community in question

Turning of dials up and down on many factors - needs to be done on a case by case basis

Many databases - but not all of them are knowledgebases - broken link to raw data or true information

Need to maintain links between samples (biobanks) and final curated data

Strategies, Challenges and Prospect of Data Management at ESO

Michael F. Sterzik, ESO

Astronomers POV

Chile-based observatory
    Most productive and lots of worldwide astronomers

Instrument line up - telescopes = microscopes (with similar cameras …)!

15 years ago were worried about data exponential growth
    not doubling every year
    working on PB scale
    Predictions were wrong - which was lucky!

SKA - square kilometer array also on ESFRI roadmap

Q: International Astronomical Union defines standards - who, why, funding, motivation (history)?

Data is still diverse - different types of data, wavelength, time (!)
Increasing volume and complexity

need an ecosystem of standards
data standards, views, interoperability, protocols
Research cycle -
idea
experiment
result
publication…

Most compute in Chile
    Metadata is main data flow until you do the big experiment; then there's a big slew of raw data, which per se are useless for analysis (volume vs value);
data processing (reduction) to create data products - quality control, master calibrations, etc.
Certified pipelines
Datasets are not acknowledged in the same way as publications - DOIs for data may be one solution.
Experiments may last five years; contractual agreement with the project to ensure updates.
Vital that publications are linked to data http://telbib.eso.org
Original data: http://archive.eso.org

Is it worth sharing data with the community?
Main message - won't get immediate ROI when you invest in creating an archive - it's a long-term investment; it's taken them years to get to about 25% of their data into the archive and it still doesn't have proper services associated with it.
Archives have to compete with exciting new projects - how do you convince your funders? It's constant struggle.

Data Management and Standardisation in Distributed Systems Biology Research

Martin Golebiewski, Heidelberg Institute for Theoretical Studies

NormSys ('Normalization and standardization for the exchange of models and data in systems biology research')

de.NBI German Network for Bioinformatics Infrastructures

Data Management for large consortium projects in systems biology:
Virtual Liver Network (Germany), ERASysAPP (ERA-Net for Systems Biology Applications)

Systems biology
Generating models from experimental data, create simulations of biological systems

Virtual liver project is a flagship german sysbio project - probably the largest in the world; dealing with patient data is a challenge (although this is national so I'm guessing the legal situation is homogeneous)

Buddy to buddy exchange of data
Don't want a data dump where data is just forgotten about
Integration tools are a major challenge but can be tackled; the social challenges are much harder. Have to convince people to use the resource; to exchange data - won't do it if they feel they're competing against each other.
Also have to cope with geographical distribution, different scientific backgrounds, managing expectations; outreach to the public…

PALS - project area liaisons - experts working on the project who collect data requirements and help train network members (same idea as the data contacts for each RI proposed by the BMB hub yesterday)

'Just enough results model' as opposed to 'just in case'
Standards are important but not enough for it to be grass roots from the community - need standardisation bodies to keep it under control; otherwise standards proliferate.

Strategies for emerging infrastructures

Babette Regierer, ISBE

ISBE - http://project.isbe.eu/

Standardisation is a key part of ISBE strategy

iverse field in systems biology, so need coordination on how best to work as a community

Stewardship - involving user communities

Support for data management, stewardship, standards from grass roots

Breakout: Understanding your RI data requirements

4 groups of 7-8; led by Cath, Rafa, Steffi & Tom
Chairs to nominate a scribe and presenter in each group

This breakout session aims to explore the data requirements of research infrastructures, how they differ with speciality, the common ground they share, and how they may change in the future.

Questions we would like you to address as a group are:

- Has your RI thought about the data it will produce?
    - This might be why you are here
    - Perhaps there could be improvements
- What challenges are there to creating a data strategy?
    - Things that might make it difficult
    - Things that you have already had problems with or overcome
- Are there missing pieces in your data strategy?
    - Areas that lack expert knowledge and planning
- How can you ensure data is a high priority for your RI?
    - People who can help raise priority
    - Actions needed to be taken

Each group will have 5 minutes to summarise the discussions

Each group will then make a short presentation to the rest of the workshop

Q - how do we deal with virtual organisation?
    Big problem - represent many orgs, plus a home org
    EU-DAT - quality assurance of service
        tradition of computing centres to be built upon
        Data seal of approval - http://www.datasealofapproval.org/en/
            Difficult process to achieve
            DANCE??? (probably DANS http://www.dans.knaw.nl/en)
        shareholders need to be aware
Q - how do you get buy in for this approval process?
    Elixir - evaluate of metrics (similar to EBI's) - need for agreement on metrics

SLA - is this needed?
    Maybe a matter to address in future

Training - Letter of understanding

Commitment to run training events, no binding requirements, but lays out what must be done by both parties.

Legal aspects of this - operational level agreements - NEED TO CAPTURE THIS

Group 1 questions

Questions we would like you to address as a group are:

- Has your RI thought about the data it will produce?
    - This might be why you are here
    - Perhaps there could be improvements
  - Data producer vs. data manager infrastructures:
    - producers (SW ELIXIR - BILS, Infrafrontier, ISBE), managers (ISBE, EBI, EGI, Infrafrontier)
    - SW-ELIXIR - BILS - does not have a strategy yet
    - Infrafrontier (mouse-human data) - Main database is the European Mouse Mutant Archive with its established policies with continuous improvement of data collection and curation processes. (Data submission come mostly from Europe. Data users are global.
    - EGI - Most of the data is with the experiments/project/VOs. Smaller part (mostly logs) are with EGI. Management strategy for this is under development.
    - ISBE - still in the beginning of the RI setup, and will start thinking about this soon.
    - EBI: Has a DMS with trying to predict data size, required compute size, proactively securing funds to procure and manage this. Distributed infrastructures complicate this further. E.g. ELIXIR UK node does not have a DMS.
    - Common points:
        - Most don't have, even those who have are incomplete and focussed on one site.
        - Important to Involve user communities in creating the strategy.

- What challenges are there to creating a data strategy?
    - Things that might make it difficult
    - Things that you have already had problems with or overcome
  - What's in it for the community?
  - Infrafrontier: Were focussed on operation and continuous improvements than strategy planning.
  - EBI: Estimating how much we need to store, compute, etc is very difficult to quantify. It's extremely difficult at the European scale.

    - BILS: Scope has not been set very well yet. What's included in life science data? Also unclear what we will store, what we won't store. Maybe directing users to public repositories wherever, whenever possible and store only the exceptional cases?
    - What you user community needs? What are the resources available for them?
    - Sharing of responsibilities and services among ELIXIR and the other BMS RIs.
    - Possible way forward:
        - Interactions among RIs not only at coordinator level.
        - Developing of national DMS and RI-level DMS in a harmonised way: Templates, workshops?

- Are there missing pieces in your data strategy?
    - Areas that lack expert knowledge and planning

- How can you ensure data strategy is a high priority for your RI?
        - People who can help raise priority
        - Actions needed to be taken
    - H2020, national calls ask for DMP - terminology.
    - Get user communities involved - can't define strategy unless you know what users need
    - define who you could delegate to

Group 3 - Discussion minutes

RI: ISBE (Babette), EATRIS (Janneke, Jeroen), ELIXIR (Peter, Rafael), EUDAT (Johannes), LCSB Luxembourg (Christophe)

- Has your RI thought about the data it will produce?
    - wrong question? maybe … data that will manage?
    - Producing data vs facilitating data management
        - data ownership?
            - EUDAT is not producing data, it provides generic services and resources for data storage, annotation, accessing, indexing and management
            - ISBE , facilitates data management
                - Should RI provides services for data production? -> Broker
                - 3 types of competences
                    - Data integration and modeling
                    - …
                    - ...
            - ELIXIR do not produce
                - facilitator of data producer
                - hosting

            - data -> knowledge
        - EATRIS do not produce data
            - is broker between industry and research
            - there is a data strategy, connecting all islands, connecting the whole pipeline
            - tools are offered to partners
    - if you provide facilities to produce data, then your infrastructure is not a producer of data
- science-centric view of RI:
    - front-end: RI allows for going from a scientific question to getting answers; such RI needs strong connection to data producers
    - back-end: dealing with a single aspect (e.g. data storage) and being part of a front-end infrastructure
- data-centric view of RI: ELIXIR takes a central part around domain-specific infrastructures and technical infrastructures

- What challenges are there to creating a data strategy?
    - ownership
    - ISBE
        - Data collection not centralized but **distributed** due to **heterogeneous data** production sites
    - EATRIS
        - data **volume** is a challenge? **privacy**, anonymization, **data sharing**, patients decide, same as in BBMRI
        - strategy about data privacy: a document is currently developed
        - quality
    - EUDAT
        - data **security** strategy: strategy under development, is on the roadmap
        - sustainability strategy: data are assets of the science societies (e.g. MPG) and must be responsible to keep those assets available; long-term strategy of scientific organisations
        - maturity of the organisation of user communities
    - ELIXIR
        - **diversity**

- Are there missing pieces in your data strategy?
    - Areas that lack expert knowledge and planning
    - ISBE
        - **harmonisation with other communities**
        - **sustainability**
            - standards development, adoption
    - EUDAT

        - **cost models** need further development
            - how to account properly for the usage of data and the usage of resources?
        - **authorisation to data**
    - EATRIS
        - **semantic web**
    - ELIXIR
        - **engagement with community**
        - standardisation !?
        - not what to do but how to do it with **limited resources**

    research vs. service in an infrastructure
    facilitation of research
    ESO: experiments done via infrastructure are made public later on

- How can you ensure data is a high priority for your RI?
    - EATRIS
        - there is a lot to do
        - projects
        - resources
    - EUDAT, ELIXIR
        - is our business

Group 1 - Discussion minutes

Breakout group 2: Brane Leskosek (ELIXIR), Jan-Willem Boiten (EATRIS), Chris Morris (INSTRUCT), Chris Evelo (EuroDISH), Tanja Ninkovic (Euro-BioImaging), Steffi Suhr (BioMedBridges), Steven Newhouse (EMBL-EBI/ELIXIR)

Key points:

- **Data needs and expertise are community-specific**
- **how can the community (or other RIs) drive change at existing RIs or repositories?**
- **what is the incentive to drive for additional funding to cover data deposition/management and sharing? Basically: is it worth making the process of setting up a new RI even more difficult by asking for more money to cover data-related needs?**
- **should RIs only drive for standardisation? very important to invest in that, rather than in storing all data - make data usable wherever they are sitting**
- **BUT: ESO example: invest in long-term data storage: over the long time, the archived data that is shared is becoming useful/is reused for new studies**

- **ELIXIR** Slowenia - considering common umbrella for several infrastructures
- ministry likes that, would be less funding for organisation and administration costs
- does merging affect data strategies?
- quantity of data that is produced e.g. in **EuroDISH** - a large part can go into ELIXIR (e.g. omics data), but extensions (additional data) needed, also requires collaborative network that works
- some EuroDISH/nutrition and food data does not have a “home”
- links are conceptually clear, but contacts are not there yet
- could one of the existing RIs create new or extend/modify existing repositories to host this data?
- still need expert annotation (food, nutrition)
- need to be able to link output from studies across different repositories
- nutrigenomics needs to be able to store complete studies
- **EATRIS**: clinical research: 90% of it is small scale in terms of data, sample size, patients - infrastructures are more scaled towards large-scale uses
- individual projects don’t have ability to deal with the specialties of the infrastructures and simply wants to get started (data in spreadsheet)
- services can produce overhead - how to do this vs. just doing it yourself
- part of the quality of deposited data is metadata etc.
- there may be projects with little incentive to share data/make data shareable
- overhead connected to data infrastructures includes: expert curators, sensitive data: safety framework, data access committees
- **Euro-BioImaging** takes care of standardisation etc., not data storage (or only a small fraction)
- political pressures e.g. to host data in a certain place or to create new centres - technical strategy needs to fit that
- scientific driver - what is worth storing?
- a given community may have a (formal) responsibility to store data, but they are not aware of research infrastructure (e.g. data produced at hospitals) - cannot take advantage of and tie into existing or developing infrastructure
- don’t overdue it with standards - lower entrance barrier
- biology data: standardise, integrate, AND link to phenotypic data

General discussion

Q: what if RI and home organisation have different or even conflicting data management policies?

Service level agreement -> service level expectation.

Breakout: Understanding the needs of data users

4 groups of 7-8; led by Cath, Rafa, Steffi & Tom
Chairs to nominate a scribe and presenter in each group

This breakout session aims to explore who will be making use of the data generated by research infrastructures, what users might need the data for and how they will use it.

Questions we would like you to address as a group are:

- Have you taken steps to understand your RI user groups?
    - Stakeholder surveys or forums
- Who are your users?
    - The people expected to interact with the RI
    - The size of these communities
    - The homogeneity of these groups
    - Other users...
- What do they need from you?
    - Service provision, long-term support
    - Training, mentoring
- What challenges can you foresee and potential actions?
    - Provision
    - Scale
    - Social
    - Discoverability

Group 1

- Have you taken steps to understand your RI user groups?
    - EATRIS: on-going
    - ELIXIR-Slovenia: not yet, counting on EXCELERATE and other ELIXIR TFs surveys
    - ELIXIR-Slovenia: interactions with users at open meetings
    - ELIXIR-Sweden: biannual open calls, constant interactions with community
    - EBI: industrial users describe needs
    - ISBE: surveys, PALs network in touch with research groups, forums, comment sections, requirement engineers, webinars
    - EGI: don't really own data, users own data, data produced elsewhere
    - MIRRI: we take these steps now, 17-18.02.2015 we looked for the possible communication with bioMedical Science, all the time we communicate with microbiology in academia

Hierarchy of service provision.

- Who are your users?
    - EATRIS: broker between research groups and industry
    - MIRRI: very diverse users from all over the world
- What do they need from you?
    - Service provision, long-term support
    - Training, mentoring
    - MIRRI: some of them need microorganisms alive, some need advice how can they communicate with these organisms
    - Co-development!
    - Iterations with users to continuously improve

- What challenges can you foresee and potential actions?
    - Using the language of the users.
    - Scaling: start-up phase -> exploitation phase, go in steps 10->100->1000->...
    - Long-tail users: training challenge.
    - Guidance without being patronizing.
    - Engage with publishers and funders.

Summary Feedback from Group 1

- Understanding Our Users
    - Surveys, Human Networks (industrial, PALs,), Annual Meetings, Webinars
    - Most RIs have a broad (non-overlapping?) user
- Interaction with Users
    - Take time to learn to speak the same language
    - Software development needs to be agile and co-developed
    - Different types of users (Power, Casual) need different things
    - Training on tools, standards, etc. important
    - Techniques may need to change as scale audience changes
    - Engaging with publishers and funders to change culture around data submission

Group 2

ELIXIR, EATRIS, EuroDISH, ISBE
Key points:

- store all data vs. channel to established, suitable public repositories
- domain specific requirements of different user groups - included not confronting people with stuff they don't need
- different communities call for different types of mapping between data
- need to lower threshold for user community to provide data in a usable way - annotations, standards, formats; provide tools to make this easy
- people in between data management system and experiment - data scientists

- how to find out data volumes coming from users - in the near and longer-term future: moving target
- need a study registry that says where the different bits of data are - similar to clinical trials registry
- ELIXIR - components that will handle vs. those that BILS can do itself
- EuroDISH - user groups are very different
- pharma - very specific requests for models that can be built into proprietary pipelines
- provide the data in a way that makes it possible for the users to get what they need/do the analyses they need to do (e.g. necessary annotations)
- goes back to question of how to drive adjustment of existing infrastructures
- identifying and combining user communities
- some part of just building something and seeing whether it works - trial and error, learning by doing
- ISBE - user surveys, currently auditing all systems biology relevant centres
- EATRIS - openclinica - TransMart: power user is evaluating whether bridges suitable
- choice of tools made based on several criteria, (1) primarily what is used by the community and (2) what is supportable/sustainable/modifiable
- patient data annotation is chaotic

Group 3

*EATRIS (...), EurobioImaging (Tanja), Infrafrontier (Sabine), LCSB Luxembourg (*Christophe?*), ELIXIR (Peter, Julie, Rafael)*

- Have you taken steps to **understand** your RI user groups?
    - EATRIS
        - No real in-depth knowledge (but good overview), some surveys
        - Biobanking surveys, but only small part of whole picture
    - EurobioImaging
        - big survey of needs for technology, data and user needs
    - Infrafrontier
        - user feedback forms
        - attend sci meetings
        - organise workshops
        - no surveys
    - ELIXIR
        - we did user personal analysis

- **Who** are your users?
    - EATRIS
        - companies (pharma), institutions, academic
        - broad group
    - Euro-BioImaging
        - biggest group is researchers
        - industry

        - high level experts in tech
    - Infrafrontier
        - scientists from different disease areas
    - ELIXIR
        - better to talk about stakeholders than users: researchers, industry, etc.

- What do they **need** from you?
    - EATRIS
        - project management
        - where expertise centres are
    - EuroBioImaging
        - access to technology and expert support, training, support with data
    - Infrafrontier
        - services around mutant mouse lines
        - tech development - freezing, shipping embryos, sperm
        - training courses
        - data on phenotyping (service and collaboration)
        - data curation, standardisation, integration
    - ELIXIR
        - data access
        - data curation
        - data integration (cross references, RDF)
        - training
        - data tools
        - compute
        - expertise

- What **challenges** can you foresee and potential **actions**?
    - EATRIS
        - how to get users on board?
        - how to reach out? more collaboration to get more impact
        - human studies approval
    - EuroBioImaging
        - service fees, grants do not foresee fees for service - **Action**: Lobby funders.
        - DM plans should include more info about how to pay for service. Joint effort with ALL the RI.
    - Infrafrontier
        - getting complete data from producers when archiving mouse lines, curate data
    - ELIXIR
        - people do not put enough effort to register resources
        - Actions: face-to-face interaction with data producers

            - provide credit/award mechanisms to provide and curate data
        - create ELIXIR community
            - engage with existing communities

Group 4

Questions we would like you to address as a group are:

- Have you taken steps to understand your RI user groups?
    - ISBE:
    - personal modeling
    - surveys
- MIRRI
    - questionnaire
    - personal contacts
    - discuss ways to use data
- EUDAT
    - questionnaire, used for f2f interviews
    - RDA driver
    - data landscape surveys
    - working groups with users
- INSTRUCT
    - service provision
    - process and use
    - User interface w/ users
    - collaborative development
- Who are your users?
    - EUDAT
        - managed repositories, repository managers from all kind of disciplines
        - from large repositories wanting to replicate data to keep stuff safe (e.g. from earth science) to the so-called long-tail data (data from small departments)
        - inhomogeneous ways to model and handle data; just for long-term archiving there is the common understanding that archived data needs to be properly packed (as AIP in terms of the OAIS archive model)
    - ALL
        - biologists
        - chemists
        - bioinformaticians
        - professional bodies
        - academics
        - public
        - students (graduate, PhD students, postdocs)

        - naive users
        - public health bodies, hospitals etc.
        - whole project consortia, communities
        - industry: SMEs, big industry, ‘7 sisters’
        - Providers of services
        - data / RI users have different knowledge levels
        - ICT

- What do they need from you?

ALL

- data
- data repositories
- data deposits (possibilities to share data within or across domain boundaries)
- instruments
- tools
- knowledge/expertise -> consulting
- samples, materials
- service products
- standards
- training
- community access
- access to news, cutting-edge knowledge
- privacy

- What challenges can you foresee and potential actions?
    - Provision
    - Scale
    - Social
    - Discoverability
    -

ALL

- INSTRUCT: Support: technicians, understanding, access, policy to apply
- MIRRI: visibility, awareness of data availability and relevance, services, tooling, software
- ALL: integration with databases, data services
- ISBE: Expert access (but to consider how to attract the experts to provide service)
- ISBE: Knowledge, specimen: not of-the-shelf solutions yet available, needs to serve the broad spectrum between beginner level until professional
- MIRRI: Expanding knowledge -> use full services available

- ALL: how to deal with the overlapping services? how to interface the services for the benefit of the user/customer
- ALL: Connections between the RIs, building workflows

Flash presentations

This section of the workshop is a chance for participants to present at the workshop; to raise topics of interest; to highlight useful research, services and experiences; or to quickly deliver an elevator pitch.

DISCUSSION

Exposure to user community

Do users care about the RIs?
    No...

What/where to submit data to?
    Wizard or tree structure
    Who to engage with?
    Service brokering

User persona matching to services/RI

Making data submission less of a pain
    Good data submission needs crediting

What is size/scale of data
    Not just the ‘long-tail’
Education
    Encouraging future users to become aware of RIs
    Make young researchers smart about RIs and data

RIs still immature
    It will take time to develop our offerings

Chris Evelo's diagram GET A PICTURE!
- Middle out
- Where does Raw Data go?
- Experiments are often more complex that RIs
- Talking with librarians

Institute level - engagement w/ community
- Low penetration
- Lack of exposure and understanding
- who is responsible for data management in the institutions?
- is there an internal policy about data management in place?

Culture change is needed in biology!
- RIs in the right place to do this training!

PhDs should be owning the data submission process much better in the future
- EBI is joining EUDAT 2020 work
- Feeding Elixir work back to EUDAT
- Look at increasing usability

Need to pay attention to social interactions
- Few true polymaths
- Equip biologists to interact with other disciplines

A big role for facilitating
- Need to know the questions that you are asking

Big need for incentives
- Publications is main driver

Data as a product of the RI
- Publication is less relevant in face of good curated data

Integrating data/combining different types
- But also organising your own data
- How to keep it
- How to retrieve it

Using generic solutions
- Good room for collaborations
- Problems from both sides
- Need to understand biologists metadata

Need to understand standards for metadata

Do you want to dip into data

Talking to board of directors - need a RI message to send, needs to be easy to understand
- Needs to be digestale
- Needs to see benefits - if it is publication then that will be what they want
- Citations of data less important
- But RI connecting data to publications

Need to lower energy barrier to storing/linking data

Incentive is getting funding

Should look to get funding for DM
- Needs of users
- Steer users towards this knowledge
- Need to be sorted out by RIs first!

Funders are defining this currently
- This needs to be the RIs driving future
- Agreements between RIs will help?
    - Sort this then support users

Dutch call - how are you going to do data management plan for this project
- Need to discrimated
- What data do you already have?
- Look at these questions and what they ask
- MRC data management plan - first call
    - Variation in understanding from users on data support
    - Few understood benefits of data access and sharing

Concluding discussion

Wrap-up of the meeting

- need culture change; scientists to feel responsible for their data
- educate and train early
- RIs role to drive cultural change in respective communities
- EUDAT2020 has training and outreach capacity and will do this work
- institutes and universities are setting up own repositories to fulfil various OA/open data requirements - now is crucial time to reach out and establish best practices (including standards, formats, annotation, etc.)
- right now funders are driving data management plans - RIs should drive this process
-

Actions

The next steps that we can take to assist RIs with data strategies and planning

- Cath - We have good ways of interacting with our users
    - Best practice on interacting with users
    - Getting engagement on data management
- Chris Evelo - Cartoons are needed
    - Data is important, so is metadata!
    - Funders/users need to understand!
    - Explain competencies
- ‘Credits’ for user as a data depositer
    - User as a research entity/authenticator
    - Needs to be better represented
    - Empowering researcher as generator/producer
- Rafa - Strategy for recognition of this
    - Credit for many types of users/RI members
    - Data strategy interest group
        - Recognition
        - Credit
        - Incentives
- Steffi - connections with e-IRG to RI advisory groups
- Babette - Interest in collaborating with industry
    - SME engagement is difficult
        - expensive for them
        - but it is very attractive for them and they want to engage
        - why RIs and not another SME

        - What makes RIs attractive to industrial users
    - ISBE - SME wouldn't come to RI for large project work
        - smaller, low-level interactions
    - Lack of knowledge from SME on open data
    - Safe data environment is needed - trust is required
        - SLA and 24/7 provision
- Users need a website with a step-wise process
    - SME after privacy of data use/submission
    - Difficult to stop Industry (of all sizes) scientists from using free stuff!
- Links between users and RIs
    - can the link feedback to RI
    - Needs to be bi-directional
- Need level of automation
    - ORCID
    - Highlight other data of interest
    - Enforcing ORCIDs is a stick
        - Need carrots! (like shows your main publications if you fill it)
- Task a group with writing best practice on data strategies
    - Not one size fits all
    - What are the common elements
    - What do you need to worry about as an RI
    - Need to move this from low-level and take to higher level factors
        - Is political/financial the starting point?
- Julie - incentives
    - Small research centres that can be targeted?
    - X-PRIZE for good data citizenry?
    - Culture change won't change overnight

Workshop report
Best practice for data strategies
Technical managers need to work on this - group to do this

Name	**Key point**	**Action**
Jeroen Beliën	THINK BIG, START SMALL, ACT NOW	solve local actual problems,be "THE BRIDGE" between users and RIs
Jan-Willem Boiten	Focus on other users than bioinformaticians; get to the real	Create simple stepwise protocols on the website how to upload the data.

	end users (the non-bioinformaticians) Get involved in data production pipelines; that's the place to grab the data. At later stage it is much harder to encourage users to deposit their data	Start with simple show cases that can be published and copied by others in the same domain.
Mikael Borg	Including cost for data management in cost of data production.	Working on ELIXIR data storage and replication strategy - advice welcome!
Cath Brooksbank	Between us the BMS-RIs have some really good ways of engaging their user communities - we should publish this.	white paper on best practice for interacting with users
Persephone Doupi		
Chris Evelo	Different levels of repositories (see flip-over) need to communicate both in software and between people (managers/funders)	
Sabine Fessele		
Andrea Giachetti		
Martin Golebiewski	- Definition of general and persistent data and metadata standards - Incentives for users to share data ('citable data')	Contact to standardization bodies (already delegate at ISO/TC 276 'biotechnology' and at German DIN) and grass-roots standardization initiatives (e.g. COMBINE)
Tom Hancocks		
Niclas Jareborg	Data strategy vs Data plan	Start working on the DS
Rafael Jimenez	Work on a strategy for recognition, credit and incentives	Create an strategy group to discuss and come up with an strategy to improve recognition, credit and incentives. Include people interested from different infrastructures (life sciences,

		e-infrastructures and other infrastructures outside life sciences).
Peter Juvan	Researchers/data producers often do not care about good practices for data management. Recognition of benefits is important, and a simple mechanism/strategy is needed.	Face-to-face interaction with data producers. Provide credit/award mechanisms for data management and curation.
Angela Krueger	Data management strategies might be overlapping between RIs and initiates	at strategic+technical level, RIs have to work together to develop common data managment strategies to use synergies (co-ordinator level needs meeting to work closer together) (data tzars)
Robert Kueffner		
Brane Leskošek	1. learn from other (outside LS) RIs, like ESO 2. use more formalised data management standards, but still simple that it would be accepted by users	1. show ESO stats as proof for funders that long-term investment is needed 2. write down simple instructions (in formal standard document) like “avoid free text”, “use structured fields only with controlled vocabularies/ontologies” and facilitate its acceptance by RIs (workshops, training?)
Julie McMurry	Synthesis research requires champions with use cases. Bolting data together for its own sake is costly and unsustainable.	Publish identifier best practice (including identifier documentation policies)
Chris Morris	Can share data policy of Instruct	
Steven Newhouse	‘RIs storing data in Elixir’ only makes sense if Elixir knows where data can be stored!	Technical Services in WP4 should establish a prtfolio of services coming from the

		nodes that can be used to store data. Consider SLAs etc.
Tanja Ninkovic	Make data submissions and instructions for annotations as simple and understandable as possible	-RIs to create user friendly portals for data submission -Include data producers (often high level experts that are supporting researchers with data collection) in training users in annotation and management
Johannes Reetz	RIs providing access to data need to be able to assign credits (in terms of the frequency of data re-usage, data citation metrics) properly on the researchers (data originators) account. For this purpose means like ORCID should be used consequently.	RIs at the forefront to the scientists must encourage them to organise their data (and metadata) in a way that makes and keeps the data meaning- and useful.
Babette Regierer	perspective of industry as user group, big industry and SMEs	start with sharing experience across RIs
Paolo Romano	there is a need for a uniform and coherent framework for semantics of all BMS RIs data, propaedeutic to an effective data integration	fill the gaps in existing minimum information models and ontologies and develop appropriate “branches” for connecting them
Gergely Sipos	Data Strategies can greatly help clarify relationships among RIs, among RI members inside each RI, among RIs and e-infrastructures therefore initiatives like BMB should continue facilitating the development of data strategies in the BMS RIs.	Check how typical data strategies look like, what topics they cover. Capture common elements. Extend those with unique features that are relevant for RIs. Reach a ‘Data Strategy template for BMS RIs’. One example of a relevant article: http://www.eiminstitute.org/library/eimi-archives/volume-1-issue-2-april-2007-edition/data-strategy
Stephanie Suhr	Institutes and universities are setting up own repositories -	RIs open up communication channels (via

	make sure data is reusable and accessible	EUDAT2020??) to provide expertise/guidance
Morris Swertz		
Christophe Trefois	RIs could be seen as too abstract and not well known in research institutes and universities	RIs should take a leader role in initiating change management in more local settings and assist such entities in changing their culture.
Janneke van Denderen	INFRASTRUCTURE WITHIN LOCAL INSTITUTES SHOULD BE IN ORDER	support the bottom topdown
Alexander Vasilenko	MIRRI-IS dataset connected to ELIXIR BioMedScience in LOD cloud	First is to think
Jan Wiebelitz		
Jarosz Yohan	Harmonise data infrastructures at the local level - Solve data integration with many data types, data sources, semantics, ontologies.	Build up common ontologies, allow “clean” data capture among our users. Increase reproducibility.

Contacts

If you are happy for other to contact you, please provide emails or LinkedIn profiles below.

Name	Infrastructure	Contact
Jeroen Beliën	EATRIS	jam.belien@vumc.nl
Jan-Willem Boiten	EATRIS	
Mikael Borg	ELIXIR Sweden	mikael.borg@bils.se
Cath Brooksbank	ELIXIR	
Persephone Doupi	PARENT	
Chris Evelo	ELIXIR Netherlands (DTL), EuroDish, dbNP, Open PHACTS	chris.evelo@maastrichtuniversity.nl
Sabine Fessele	INFRAFRONTIER	sabine.fessele@helmholtz-muenchen.de
Andrea Giachetti	Instruct	giachetti@cerm.unifi.it
Martin Golebiewski	ISBE, COMBINE, de.NBI, ERASysAPP (FAIRdom)	martin.golebiewski@h-its.org
Tom Hancocks	BioMedBridges	tom@ebi.ac.uk LinkedIn
Niclas Jareborg	ELIXIR Sweden	niclas.jareborg@bils.se
Rafael Jimenez	ELIXIR	rafael.jimenez@elixir-europe.org
Peter Juvan	Elixir Slovenia	peter.juvan@mf.uni-lj.si
Antje Keppler	Euro-BioImaging	
Angela Krueger	ISBE	angela.krueger@mdc-berlin.de
Robert Kueffner	Helmholtz Center Munich	
Brane Leskošek	ELIXIR Slovenia	brane.leskosek@mf.uni-lj.si
Julie McMurry	ELIXIR	jmcmurry@ebi.ac.uk
Chris Morris	Instruct	chris.morris@stfc.ac.uk

Steven Newhouse	ELIXIR	steven.newhouse@ebi.ac.uk
Tanja Ninkovic	Euro-BioImaging	
Johannes Reetz	EUDAT	reetz@rzg.mpg.de
Babette Regierer	ISBE	bregierer@gmail.com
Paolo Romano	MIRRI	paolo.romano@hsanmartino.it http://it.linkedin.com/in/paoloromano
Gergely Sipos	EGI	
Stephanie Suhr	BioMedBridges	
Morris Swertz	BBMRI	
Christophe Trefois	University of Luxembourg - LCSB	christophe.trefois@uni.lu
Janneke van Denderen	EATRIS	
Alexander Vasilenko	MIRRI	vanvkm@gmail.com
Jan Wiebelitz	e-IRGSP4	wiebelitz@dcsec.uni-hannover.de
Jarosz Yohan	University of Luxembourg - LCSB	yohan.jarosz@uni.lu