E-Infrastructure support for the life sciences: Preparing for the data deluge

A BioMedBridges knowledge exchange workshop hosted by ELIXIR

This document has now been closed for editing (status 23 May 2014). A summary of the workshop will be circulated shortly.

Welcome to the 'Living Document' for this BioMedBridges workshop. It was used to take notes on presentations, record discussions and create a collection of challenges and answers covered during the workshop. Workshop participants have contributed, added in links and resources and made corrections where needed.

Slide presentations from the workshop are available on the BMB webpage.

Participants

Name	**Affiliation**
Bernardi, Sergio	PRACE
Blomberg, Niklas	ELIXIR
Boiten, Jan-Willem	CTMM/EATRIS
Borg, Mikael	BILS/ELIXIR
Brooks, Tim	Public Health England (PHE)/ERINHA
Butcher, Sarah	Imperial College London/ISBE
Capone, Vincenzo	DANTE

Cochrane, Guy	EMBL-EBI/ELIXIR
Cook, Charles	EMBL-EBI/EMBRC
Corpas, Manuel	TGAC/ELIXIR
Di Meglio, Alberto	CERN
Ferrari, Tiziana	EGI
Geddes, Neil	STFC
Goble, Carole	University of Manchester/ISBE
Hancocks, Tom	EMBL-EBI/BioMedBridges
Henderson, Tamsin	DANTE
Hermjakob, Henning	EMBL-EBI/ELIXIR
Hughes-Jones, Richard	DANTE
Jimenez Lozano, Natalia	Bull
Jimenez, Rafael	ELIXIR
Lengert, Wolfgang	European Space Agency
Maurice, Paul	DANTE
Minaricova, Maria	DANTE
Morris, Chris	STFC/INSTRUCT
Neerincx, Pieter	UMCG/BBMRI
Newhouse, Steven	EMBL-EBI/ELIXIR
Oster, Per	CSC/EUDAT
Pietro Maggi, Giorgio	INFN
Roeth, Gunter	Bull
Sipos, Gergely	EGI
Stanford, Natalie	University of Manchester/ISBE
Suhr, Stephanie	EMBL-EBI/BioMedBridges

Swedlow, Jason	University of Dundee/Euro-BioImaging
Szomoru, Arpad	JIVE
Trimming, Matthew	Maxxim Consulting

European Infrastructures

Biomedical Science Infrastructures

- BBMRI
- EATRIS
- ECRIN
- Elixir
- EMBRC
- ERINHA
- EU-OPENSCREEN
- Euro-BioImaging
- Infrafrontier
- Instruct
- ISBE
- MIRRI

e-Infrastructures

- EGI
- GÉANT
- DANTE
- PRACE
- EUDAT
- WLCG

1 Challenges of big data

Rafael Jimenez

If you have any questions or comments please write them below

QUESTION (EGI): is the data for depositing for permanent archiving and curation? solutions are developed for data preservation in earth science (Alliance for Permanent Archives)
PREPARING

Question (CERN): what does Compute include?
Is numerical processing? Data analytics? Combining different types of data, visualization, etc.?

> Is storage really a problem? Preservation, transfer and use are bigger problems
>
> Storage requirements is increasing faster than are producing. 18 mths doubling time for storage. Disk density more than IO or fall in price is the problem.
>
> Tech not a problem, cost is the problem, but comes back to new tech solutions to try and solve this.
>
> 100s of file formats that are not suitable for vast data volumes

Compared with physics, biological data is much greater

> A: I would say it is greater in variety, not sure about size. This is why I asked how much of a technological problem this is (disk companies are coming out with new types of disk today that can store many times the current size, many TB). If every data producer must store locally, it is probably more a cost problem, not technological. Having centralized storage might help bringing down the cost in addition to allowing easier data management.

Is question how to provision storage independent of biology?

Need to get ahead of the community of data generators and users
Users access data in a different way

IO is key factor in biology - profile of access to databases shows that 40%? of data is access soon after submission, repeat access

life sciences: lots of comparisons of data resources with other data resources, including big with big

Raw data storage less a problem in physics as it is processed and then discarded

Particle physics is a closed set of people who access

Genomics has a rich approach to access, not of raw data, but annotated and integrated

Imaging is perhaps the big problem as compression is much more difficult
- Standardisation - similar to geo-spatial world
- Digital pathology
  - data producers compress anyway
  - doctors only get compressed files

Research and clinical care are hand in hand with the new techs
- Keeping data close (and secure) to patients

Local procurement in hospitals - commercial approaches?
- Distributed data then
- But potential need to share data between many locations in future
- Connectivity then becomes key
- Turn-around times important for patients
- Shorten this in future

Performance demands local data storage and access
- More difficulties for data integration

Beacon project
- https://genomicsandhealth.org/our-work/working-groups/data-working-group
- Global Alliance for Genomics & Health
- Collection of genomics
- Provide info on genomes with a particular variant

- Global Alliance for Genetics in Health
- Question of anonymisation of data at this level
- Dispersed and distinct ethics committees for access to data

Question (CERN, I put it here for a generic discussion, but applies to all presentations [Note: this seems to be repeated, probably because I put here a question, but also someone added notes of the discussion]): Is Storage really such a big challenge? In what sense? Capacity grows continuously. Or is it more question of how to use and preserve the data?

> It clearly is; data storage needs are more rapidly growing than the costs for storage are coming down. So, this growth is not sustainable in its current form.
> The rate of data aquisition by some instruments is increasing more quickly that network bandwidth or disk capacity: notably sequencers and Pilatus detectors.

1.1 Science community data challenges

1.1.1 Genomics

Pieter Neerincx (UMCG, BBMRI)

PROBLEMS: anonymization of private data. Solution: no sharing of data about individuals, only about groups (aggregated information for groups of information). Code is transferred to the data

> I don't see why DNA is different than any other medical data. In itself it is still anonymous unless you have a sample to compare to, but that is basically the same for any other medical data as well. (Jan-Willem)
>
> See: *Science* 18 January 2013:
> Vol. 339 no. 6117 p. 262
> DOI: 10.1126/science.339.6117.262
> NEWS & ANALYSIS GENETICS
> Genealogy Databases Enable Naming of Anonymous DNA Donors
> John Bohannon

Question: Several projects are moving to build cloud-based resources for disease-related genomics data (e.g., http://oicr.on.ca/report/international-cancer-genome-consortium). What's BMB's relationship with these growing projects?

> Security is a challenge, but is it a big data challenge? AAI is a big topic maybe should be discussed in another meeting. BMB is planning to organize this as well.

QUESTIONS (DANTE)
Question: What type of IO bound - disk or network?

> Both, but when it is network it is usually the connection to shared storage. Most applications are embarrassingly parallel, so there is no network bottleneck between servers to compute.

How big are the data transfers?

Varies. Could be from 10-100Mb files of variant calls to an entire genome of raw FASTQ reads of the size of ~100Gb or potentially many whole genomes into the Tb size

Is there a real-time requirement?

No; though for clinical applications speed is very important. Currently from whole genome sequencing to diagnosis in < 7 days. We aim to get that down to < 2 days before the end of this year.

Question (EGI): shipping of computing to data requires availability of local compute facilities, how are these federated? how is data discovered?

TO ANSWER

1.1.2 Proteomics

Henning Hermjakob (EMBL-EBI, ELIXIR)

ISSUES. Diversity of data, only metadata is shared

Question: PRIDE is one of many proteomics repos-- any hope of linkage, rationalisation, etc.?

HH: Most of what I showed relates to PRIDE as part of the international ProteomeXchange consortium, which does exactly that, link proteomics repositories through a central metadata format and searchable index:
Vizcaíno JA, et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol. 2014 Mar 10;32(3):223-6. doi:10.1038/nbt.2839

Which characteristic(s) of the big data problem appear in this use case? Total size?

Challenge transferring files (upload/download). Using Aspera (Commercial) to transfer files faster (x10 times)

What are the prerequisites for using Aspera? Is it something we can use for any data transfer against cost-effective conditions?

HH: The Aspera server licence costs (significant) money. Client use is free. As described by Guy, there are efforts for open alternatives.

1.1.3 Imaging

Jason Swedlow (U Dundee, Euro-BioImaging)
Cloud resources for computing needed
Open source project for services to be delivered to users
Volume: 10-100 TB per users per dataset, common data repository needed

SOLUTION: cloud infrastructure co-located to data for data mining etc.

Just few sites producing data?

> Sorry for moving a quickly. There are a few hundred to 1000 imaging facilities in the EU. As I mentioned, Euro-BioImaging has defined specific technologies, and methods for identifying new up and coming methodologies that are appropriate to run as an open access resource. Those are the basis of the EoI Nodes we now have. The modeling in the PPT is based on these technologies

Cloud services for computing. More details? Bring tools (VM) to data hosted in a centralized repository.

> Admittedly, this is still being defined. EMBL'S Embassy is an example, but several exist

Question (EGI): will the data repository be distributed rather than just centralized at EBI?

> Initially, we think its wise to start with a sort of proof of concept, thus our focus on rerefence images, linked to data resources that are already well established, e.g., linking phenotypic image-based screens to genoic resources. So in the short-term, having a small set of defined repositories may make sense, especially where the first goal is to provide linkage with molecular resources. In the longer-term, distributed resources are certainly possible.

PROPOSAL: having a centralized catalogue instead of a single (big as you like) system. There are issues about coherence of the data between the various sources, but nothing that hasn't been seen before in other fields (e.g. LHC ATLAS distributed data systems).. (Enzo Capone - DANTE)

1.1.4 Metabolomics

Natalie Stanford (U Manchester, ISBE)

Raw data: 15 MB for each sample
Quality assurance and checking to verify quality of the machine readings
Metadata needed to reuse of information, to provide information about the samples, how collected
Volatility: data gets better with technology, in 2 years cycle data quality information improves, short term cycle

If you have any questions or comments please write them below

Question (EGI): compute requirements of the data?

Analysis on .txt file is fairly fast on standard computers (or at least what we do). It will also depend on the size of the data sets.

Question (EUDAT): Why is not the data kept in net.cdf format as long as possible? It is a standardised format for packing data and many applications and libraries exists to for un-pack or extract data from net.cdf files?

I suspect this is changed early in Manchester due to the scripts they use being generated in house to use on .txt files. There are a number of applications and libraries to use it in raw format. I'm not sure if this affects the computing time though.

Metabolomics often uses NMR as well as mass spec and these data have different formats, rates of generation and sizes to the proteomics data. may have to be analysed together with the mass spec as part of the same study. Raw Data size depend very much on the assays run per sample. For instance, can have datasets of a few GB per assay but more complex assays yield several hundred GB. This can result in Petabyte (PB) outputs per year for a large facility. There are extensive pre-processing steps, often requiring cross-referencing of datasets, local QC standards.

Estimates of numbers of sites generating metabolomics data in 5 years time across Europe are expected to be inaccurate:
large facilities ~5
all other producers ~100
recent ISBE survey identified that ~ a third of systems biology respondents expected to use some type of metabolomics in their research in the future (D9.1)

1.1.5 Clinical data

Jan-Willem Boiten (CTMM, EATRIS)

EATIRS shares infrastructure with bioimaging and others RIs
Security and trust in archiving data is key
Interface and usability of tools is important
Compute needs: computation needed for image processing, DNA/RNA sequencing processing pipelines

1.2 Data fluidity

Guy Cochrane

Storage and processing of information with large volumes become a challenge
No centralized data sets, but small instruments spread around the world producing data in ad hoc way → DISPERSED SCIENCE, data being picked up remotely in the world and downloaded
Data needs to be aggregated for analysis

Data growth. EBI databases. Exponential growth. 12 months doubling time. 5-10 PB.
How do you replicate 10PB?

GC - The current model is to replicate at write time - works quickly as we have substantial network connectivity between data centres.

Number of users life science data vs. other disciplines? access requests to databases, I/O data enrichment/curation needs

GC - ELIXIR surveyed this in detail - Steffi?

Example of how much cost to sequence taking into account the whole process (sampling, manipulation/experiment, etc.) - cost for sequencing very low, surrounding costs can be (very) high, conditions under which sample is taken/sample unique
VS storage cost taking into account years of storage, electricity, IT people?
In practice you often cannot regenerate the data, because the sample is depleted. For patient data you cannot regenerate the sample when disease has progressed in the meantime.

Some sample data not reproducible. OK. But how often this happens? Is this the trend? -> depends on (scientific) context

- Efficiency:
    - Storage: compression.
        - Models are needed to ensure that you minimise lossiness while achieving the desired economies. Compression itself can be expensive.
    - Transport: Protocols
    - Partitioning: Make sure that you store and dispatch the data in a way that best addresses the user needs (maintain biological context)

Can we use CRAM for other data that is not sequence data? Degree of precision is inversely related to degree of compression.

GC - not directly as it uses sequence data-specific characteristics. However, the reference and variation from reference basis is generic - taking very much from image and video compression approaches.

What about standard compression models/tools? VS CRAM.

GC - CRAM uses well established algorithms (eg. RLE, Golomb-Rice, etc.) and structures sequence data appropriately to leverage these.

Question: (JWB): can we increase compression at a later stage in the study: so, initially store in lossless mode, but later on when archiving increase the compression?

GC - Yes - although we don't have anything to support this in the code. We would like to provide a way of streaming from a CRAM file an output at any level of lossy compression below that provided in the source data. Making this an edit on the source data would also be an option, but again no code for this yet. Think about the timelines, though - there is no enormous value to increasing compression of a data set that is from a year or more in the past as the cost of maintaining this data set will be greatly reduced - this is one of the odd features of exponential growth of disk capacity per unit cost - the legacy data don't matter in volume terms nearly as much as the future data.

Why Proteomics use Aspera and Genomics UDT? What is the difference?
GC - We still use Aspera in genomics - we are developing an alternative a) to add functionality that can't be added in Aspera and b) to avoid having a system for which paid server licences are required.

Data partitioning. Do e-infrastructures have solutions for this? Can we split into chunks all the life science data?

GC - In genomics there is a clear reference genome model. In environmental sequencing, the indices may be functions (around gene/pathway) or taxonomic (around taxa and clades).

Comment (Pieter): the partitioning will help a lot for follow up / down stream analysis of large data sets once the data is available from a large community repository, but unfortunately does not help much if you created a large data set and want to upload it to the community repository.

GC - true, but freeing up network and IO bandwidth in general will create more space for these submissions (sure - depends how well joined up the services are). Also, one would hope that there is are many data retrievals from repositories for each deposition.

Are there algorithms that you can apply to data access history in order to help inform the most appropriate structure.

GC - We can tell which data sets are retrieved, but not how they are used and - because there's no real partitioning service in place yet - we don't know how much is disgarded. In time, we will be able to track this at least from EBI's repositories.

QUESTION (EGI): data is distributed but volume is a major issue. How is data consolidated and computed? how it is accessed if volume makes downloading unfeasible? Does a federated cloud infrastructure match the needs to ensure that large data does not need to be centralized and can be kept distributed?

GC - EBI and ELIXIR are piloting cloud services ('Embassy' project) provided from compute co-located with data storage. We can mount data directly onto these instances for direct compute. Also, we just launched (yesterday) an EGA (controlled access data) service from CRG Barcelona - in part to bring data closer to compute in e.g. the Barcelona Supercomputer Centre.

COMMENt: Cloud computing takes the geography out of the equation: that is usually not allowed for patient data. One needs to know in which country the data is stored for legal reasons: the country needs to have adequate privacy protection laws. [There are ways to put the geography back into the data using a "cloud embassy"]

- GC - This is one of the interesting discussions that I think will be pushed forward through the **Global Alliance for Genomics and Health initiative**. While some legislations will prevent any non-domestic data export, some level of summary/aggregate data may become acceptable. If legislations allow, a cloud Embassy where a nation has ownership, control and privacy despite being on foreign territory is certainly technically available.
    - This might be an issue that is also influenced by soft factors, such as "public opinion/perception" (this may certainly be true e.g. for Germany, from experience also from other issues). So safer not to plan on embassy clouds elsewhere??

1.3 e-Infrastructures

1.3.1 EGI

Tiziana Ferrari (EGI)

EGI looking at procurement support for user storage needs - what is the time frame?

TF: we will study the legal requirements for this activity starting in 2015. We also need to understand what is the interest of RIs in this activity: would European-scale procurement be something the RIs want to look at for having a European level coordination?

In other disciplines a different approach has been chosen, that is to move data to the computation, as the price of connectivity decreases at 2 orders of magnitude faster than price of storage. Why do you think to overturn this criterium? Do you have a different appreciation of the storage vs. network price ratio? (DANTE)

TF: for cloud specifically, hosting big data in one cloud infrastructure has major cost and efficiency issues and choosing one provider causes vendor lock-in. With a federated cloud, distributed computing can be co-located with data archives, so that compute platforms can be brokered where data is located

Comment: Data transfer via gridFTP, http, or WebDAV are blockers for imaging.

Question to the comment: Is that remote visualization, i.e. remote interactive manipulation of large data?

TF: yes these are interfaces and protocols in the Grid platform that need to be supported for data retrieval and ingestion. For the cloud platform, the community can deploy the services exposing custom protocols and interfaces, so the cloud paradigm adds extra flexibility

QUESTIONS (from EGI to RIs):
- RI and e-Infrastructure sustainability, what synergies?
- Coordinated procurement of resources/services at European scale?
- Sharing of costs of operations?
- Common backbone of ICT services for RIs with EC support for procurement and operations?

Question (Alberto): we need to really commoditize lower layer services. Does it make sense to keep running “research institute operated“ infrastructure services? Focus on higher level sevices should be the next step.

TF: both levels of service can be profitably provided. For commodity services, there is still a need for European coordination of service procurement for RIs. Procurement will allow the participation of commercial users. Data will remain distributed posing computational challenges, it is to be demonstrated how commercial cloud offers can address this. The EGI vision is the hybrid model: institutional clouds integrated by commercial ones.

1.3.2 EUDAT

Per Öster (CSC)

- EUDAT project ends in Dec 2015? Sustainable? Follow-up proposal (bigger?) in preparation - what happens to data in B2xyz services if this does not get funded?
    - The B2SHARE service is professionally managed, the data is stored at Datacenter CSC Kajaani, certified by the ISO/IEC 27001:2005 standard for its information security management system. For the free B2SHARE service, data

and metadata is committed to be downloadable for two years in case the upload service has to be closed.

- B2Sshare for sharing/uploading data. Difference to figShare?
  - Functionality is similar(?)
  - Concerning policy and business model B2SHARE is European, non-profit and no paid services (yet!). *Paid service levels are under development*.
- B2find: metadata catalogue. Metadata from where?
  - Any organisation that allow or want to have their metadata harvested. Present sources is easiest invetigated at http://b2find.eudat.eu/dataset.
- Storage of just files? LS most useful data are in structured databases not files.
  - It is a data sharing service but with the limitation that it is datasets (files) tahat are shared. A DB service is not included but if you can pack you DB as a file it is fine to share.
- B2safe/B2stage interesting to connect data to HPC and EGI computational resources. If we use EUDAT technologies will it be much easier to move data to HPC/EGI facilities? Yes, that is one of the basic ideas.
  What does the B2xyz user community look like - long tail vs. big customers? What scientific disciplines?
  Communities involved in the present EUDAT project are listed at: http://eudat.eu/eudat-communities

1.3.3 GÉANT

Richard Hughes-Jones (DANTE)

GEANT provides a high-bandwidth, high-performance pan-European communications infrastructure serving Europe's research and education community.
It does that by interconnecting the European NRENs using point of presence in each country.
It also provides Global access for the end users & researchers.

GEANT provides the following services:

- Connectivity
  - IP, up to 100Gbps access
  - MD-VPNs (L2 and L3)
  - BoD and Point-to-point circuits
  - Wavelengths: 10-100Gbps
- End to end Performance
  - perfSONAR – *Network monitoring & testing*
  - eduPERT – *Performance troubleshooting*
- AAI: eduGAIN – *Secure access, single sign-on*

- One Stop Shop
  - Consultancy
  - International co-ordination
  - Bespoke solutions

GEANT is ready for the data deluge:

- GEANT Optical transmission platform provides 500Gbps super-channels
- Which are easy to upgrade to even higher bandwidths
- Built on top of that is a routing infrastructure that provide up to 100Gbps links to our customers
- GEANT and the NRENs engineer their networks to provide sufficient head room to allow the users to perform high bandwidth data transfers
- The GEANT community believe that our backbone infrastructure is well prepared to satisfy future user networking requirements

Supporting the Life Science Community

GÉANT and NRENs are committed to supporting global Science and Big Data and the Life Sciences is a priority for GÉANT.
Examples of our work with other projects may be found at
http://www.geant.net/Users/Pages/home.aspx

Questions

- Data could be produced faster than transferring Bottleneck? Looks affordable to upgrade networks. Looks cheaper than maintaining storage.

Google fiber. 1Gbits connection at home for less than £30? How is this possible? We need this! Can GEANT provide this? Is this technology available to network providers (not google)?

EC: before I elaborate on this, a caveat is needed: GEANT doesn't provide access services, it's the NREN in every single country that does this, and as we've shown in our presentation the price and the policies are different among the countries. That said, Google has clearly a different business model and also, more important, mission, that is to provide content; in this respect, the Google Fiber pricing is not thought to generate revenues out of it, but to be complementary to the actual company mission - and also, it is sold only in some US cities.. Another point I'd like to stress is that almost every commercial network is calibrated for a 95% of usage of its backbone, while in the GEANT+NREN community this number is 50%, that means that if the average usage of a network segment exceeds ½ its capacity for a sufficient amount of time, than the capacity is permanently doubled. This means that whatever access bandwidth one buys from a commercial provider (unless it's a specific and expensive service with dedicated bandwidth), the real speed

in accessing the global Internet will be a fraction of that nominal access speed (1Gb access turns to download at 50Mb). This does not happen with GEANT, where the access speed that a user pays for is usually completely available to him (1Gb access means to download at 1Gb).

1.3.4 PRACE

Sergio Bernardi (PRACE)

- access model based solely on scientific excellence: access is based on peer-review carried out by a panel of reviewers experts in the reasearch area of the proposal there is no formula that provide proposals coming from researchers of the HM states with any advantage (see link http://prace-ri.eu/Peer-Review)
- time from proposal submission to decision to access: the decision and award of resources, including the answers from the applicants, takes usually 4-5 months actual start of work for the users depends on the centers.
- any restrictions re: size of project (HPC needs)? Answer: there is a parallel technical assessment on the feasibility of the proposal from the technical point of view..if proposal requests (or involves) resources that go beyond the actual capacity of the selected system then the proposal is flagged problematic and may not go through.
- open question whether services can remain free at point of use in the future - alternative options? Answer: let me clarify my statement that I realized was somehow misleading.. the computing services will continue to be free at point of use for scientists. What may change is that for some kind of users (industry, specialized reserved access like urgent computing..) access may be regulated by specific agreements between PRACE and the entities that require it. The PRACE funding model is under discussion in order to guarantee sustainability in the long terms beyond 2015.
- How (and when - decisions on access, see above) do I transfer my 7 multi-TB datasets to PRACE for processing? Answer: once the access has been granted the user will refer to the center that is taking care of all aspect that concerns the data needed for the computation. The center gets in touch with the awarded users in order to plan for any transfer needed.

1.3.5 CERN/LHC

Alberto Di Meglio (CERN)

The LHC is a long-term project lasting several decades from conception to decommissioning. It is organized in alternate phases of operations and shut-down periods, during which the machine is repaired and upgraded. In the next 10 years the data rates out of the detectors will increase many times, up to 50 times according to some models.

CERN is aware that the computing and data requirements, which were sort of expclusive to LHC at the beginning, are increasingly more and more common to many other scientific disciplines. A document describing a vision for E-Infrastructures in the 21st century has been published by the

EIROForum members (CERN, EFDA, EMBL, ESA, ESO, ESRF, ILL, XFEL). The main concept is the definition of a hybrid public-private model where infrastructure and services can be provided by collaborations of research centres and commercial companies. The concept is implemented as a network of Research Accelerator Hubs (ReAcH) of which CERN and EMBL-EBI would be the first two prototypes in 2014.

Several initiatives are being defined or upgraded to support this model. Helix-Nebula is an EC FP7 funded projects and a consortium of research labs (CERN, EMBL, ESA, PIC) and commercial service providers to test ral science use cases on commercial and public cloud infrastructures. It has just lauch the HN Marketplace as a place where services can be offered and bought using different business models (pay-per-use, public funding, etc.)

The CERN openlab is a public-private partnership between CERN and several major IT companies to run joint research projects on IT technology for future use in LHC. It is now being expanded to involve more research centres (including EMBL-EBI) and major international projects. The goal is to identify and work together on common IT challenges across different scientific domains. Six major areas of work have been identified and the results have been published in a whitepaper written in collaboration with the EIROForum members and the LHC experiments (https://zenodo.org/record/8765). It will be the basis for the work to be done in CERN openlab Phase V in the next three years.

Questions:

- Storing ca. 5 PB per day - type of data (raw? processed? time frame for storage - rolling storage/data dismissed after a set period?)
  5 PB/day is what will come out from the initial filtering and reconstruction process and augmented with simulation data. It's a mix of raw and processed data. It must be stored and preserved for as long as possible. At least this is the current strategy. Should this prove too difficult or onerous in the future, some other strategy might be devised.
- Currently 100PB
  This is the total amount of data (raw, simulated and processed) stored after three years of operations
- what is the relationship between EGI and Helix Nebula?
  EGI.eu is a partner in HN. The EGI FedCloud is one of the cloud infrastructures connected to the HN blue box broker
- How will the two prototype research accelerator hubs exactly look like, what kind of services and for who will they provide?
  This is being discussed
- Cloud providers that want to work with LS might need reference data. Are they planning to replicate current repositories of data?

    - I cannot answer this question without knowing more about how data is used. However I doubt cloud providers would want to replicate existing data repositories. However, such repositories might get suitable interfaces that allow them to be accessed as cloud services
- It would be lovely to understand how the model of a single data producer distributing datasets to a defined community maps to a distributed data generator and data consumer problem.
    - WLCG (the LHC infrastructure) is not really a single data producer. The initial data is certainly produced at CERN, but it is distributed for analysis and simulations across hundreds of institutes that produce more data. Moreover much of the processed data has to be shared back. The community is defined, but quite large and the sites using and producing data go from very large centres (Tier 1 and 2) to single university departments (Tier 3). Actually the need to share information across members of the same Tier without going back up to Tier 2, 1 or even 0 is prompting a revision of the original Monarc model to allow more flexibility. This migt be based on the concept of federated data storage using standard protocols (as HTTP) or tools from EUDAT.

1.3 Major challenges identified (round-up of challenges)

If you have any questions or comments please write them below

1.4 Open discussion

If you have any questions or comments please write them below

1.5 Science community use cases (group sessions)

The size of these groups will depend on the number of participants

1.5.1 Genomics

1.5.2 Proteomics

PDBe validation - does not valid raw data
Not credible to validate raw data
Potential loss of good data in

1.5.3 Imaging

Images not post-its
Imaging is getting big because of smartphone cameras!
Stakeholders
Technology

Euro-BioImaging

Carving up the problem

100-1000 sites producing images

Includes hospitals, industry and research

Only community resources that can be shared

5 years - 50 sites?

1.5PB/y - 7PB/y in 5 years

2-5% of all imaging data collected in Europe!

15% (~1PB) of this data is useful to be shared

Rest remains with data generator and their research

Archiving - nothing on horizon

Need for reference database/resource

Compute - local by generator/user

Some nodes may have expertise/compute dedicated to this

Associated with the repository at the imaging facility - cloud or local

Resources already exist for compute - but independent of imaging sites

Lack of communication!

Digital pathology 2TB/day - €250k a machine, growing number

1.5.4 Metabolomics

- data production is changing

5 years - raw data from experiments 5PB

Processed at GB level

Generic users make use of lots of resources to re-analyse community data

3 clouds - delivery of tools and data to users

Restrictions on movement of data

Legal/ethical problems of sharing

Generic cloud - different data brought together for basic analysis

Commercial handlers of data

Data standards from industry and community

Quantification against datasets from different sources/locations

Means analysis is against different requirements and challenges

Varying level of standards from different tech platforms

Software varies a great deal, different file formats

FOLLOW-UP WITH SARAH on standards for metabolomics

Comparable datasets from different machines/tech?
Varied way of generating data for metabolomics

Question - is there a European standards agency
ETSI? Experience in electronics and computing
Open-grid forums have led the way with computing too, but not specific
ISO not always the solution
Genomics is very much a grass roots standards

Would the Open Grid Forum be a suitable place to discuss this world wide? It has a current focus on GRID and Bandwidth on Demand called NSI, but it is also referenced by cloud style distributed computing. This might not be appropriate if “standardisation” is used to mean calibration of the instruments.

- in a number of countries, very large groups are starting to set up national resources that will have 20-30 mass pecs and NMR spectroscopy (currently smaller numbers of machines, either/or)
- e.g. UK about 20 universities doing metabolomics, but also national centre with >20 machines in a few years
- UK national centre expecting to produce ca. 6 PB of raw data over the next 2-3 years
- currently growing about 1 PB/year, in five years might be 2-3 PB/year
- processed data is text files
- raw data is not dismissed as new results can be obtained a few months later
- data processing is not a standardised - different groups have different ways of processing data; this is a big problem for metabolomics
- currently a lot of data ends up in small, experiment-specific data repositories
- ideally will go into metabolites in future
- might push data into repository and restrict access to certain users
- where data production and processing happens may differ on data for research or clinical use
- in reality, no one size fits all - there will still be requirements also e.g. for local storage, processing etc.

factors:

- funding
- legal/ethical framework

1.5.5 Clinical data

Lots already covered with genomics in this area

Data production - lots of data produced for clinical care not research
- Nat lang processing

Patient records/reports/demographics
- Raw/Processed data

Disease progression
- Long term collection of data
- Difficult to align patients with similar conditions

5 years time
- 10-100GB of data per patient
- Genome sequences? Exomes?

Specialist user interfaces for clinicians
- Not just specialist

Trends
- Patient stratification
- Individual characteristics

Data availability
- Berlin Wall between clinical care and research
- Big barriers to break down
  - Risk averse to data sharing
  - Possessive of research data
    - Keep on premises and in control of data
  - Politics is a big driving factor!
    - Litigation a big fear

IT side
- Need to build trust - very welcome!
- IP protection
- Interoperability and normalisation
  - Each hospital needs to work on this

Insurance companies

- Like data
- Will sponsor
- Sharing is not a big thing for general public

General Practice not Hospital is best place to approach clinical data

Increase in personal monitoring devices

- Patient to become more central in their own healthcare
- Tech and social drivers

Reference data

Factors

Science - Reproducibility, uniquness of samples, processed/raw data
Financial - costs of everything!
Technical - IT, instrument, biology/chemistry/physics
Political - geographic, industry/academic spheres
Social - public support/opposition to data sharing
ELSI - Ethical Legal Social Issues
“Language” - translation into a common (English?) language for large-scale research

Others

Changing scientists habits in how they do science
- Standards, best practices

How funding is used
- Can you buy services instead of investing in infrastructure?
- Gov may prefer investment in capital or employment
- Planning of services not really part of scientific funding at the current time
  - Will is against procurement due to shifting factors of access and cost
  - Can this be changed/influenced?
  - Is this something that could be lobbied for?
- Refreshment of assets every 36 months?
- European level of procurement
- Attention to side effects of buying services - VAT etc, very difficult
  - Having assets gives more flexibility than using a service
- Not truly flexible - grow it? use it?
  - with a service you can increase as and when needed

Trust of researchers using infrastructures - do they trust us to build the right services for them?
They might prefer building their own, instead of buying a service/using an infrastructure
- Infrastructure being cost-effective?
- Need community to request better e-infrastructure
- Data deposition and metadata requirements doesn’t help data sharing

Involve user community - still not there after many years!

Can’t let users drive all the time, need to show the way for them to follow at time
- Interactions at every level are needed

Building up of private cloud services to avoid commercial offerings
German data cannot be held outside of Germany etc

coordination of efforts (e.g. procurement) between RIs and e-infrastructures on European level
-> powerful, economies of scale

2.1 Solutions for big data

2.1.1 Earth satellite data

Wolfgang Lengert (ESA)

If you have any questions or comments please write them below

Earth Observation - not really big data, just lots of small ones

Satellite is the instrument - reception, processing and dissemination to research centres

Commercial spin offs

Monitoring of Earth, 9 societal benefit areas - focus on geo hazards
Earthquakes and volcano research

Working with CERN and EMBL on Helix Nebula

Require infrastructure, multi tennant provider = big pool for data supply/use

Open data policy with space agencies
- Make data open to uni/research
- Move towards a generic infrastructure

Ground displacement satellite monitoring
- Lots of applications
- Earthquake/volcano/subsidence
- 5m resolution over course of a year

Science and commercial use
- IPR in new impacts and offerings for many areas
- Monitoring of rainfall - lots of uses farming/engineering/climate/weather modelling

VM desktop with suite of tools
- Single sign on
- Working with DANTE
- Civil engineering collabroations to design earthquake house plans in prone areas
- VM by Helix Nebula

Google Earth Engine - competition, use of data by Google too

Not involved to keep data open access

Long-term goal

Q Jason: how many instruments?

radar, sea-surface temperature, ocean colour, gravity, electromagnetic, etc.

Sentinel satellite networks

Linking this geo data with biological environment data

Compute sites

Networking

Data

3TB of 20 years
2TB atmos
80TB sea temp level data
PB of oceanographic data
Sentinel = 2TB of data a day
Need a infrastructure to deal with this
Google keen to get involved!
Helix Nebula is a way to retain control over infra and data
Cloud enables science to happen, but isn't the whole part of Helix Nebula
Tech drives science in this case

2.1.2 Radio astronomy data

Arpad Szomoru (JIVE)

Data transport for radio astronomy

VLBI Network - combine radio telescopes 70 days a year as one instrument
JIVE oversees this collaboration and brings data togehter
To become and ERIC

cm wavelength of EM spectrum

Use it to explore stellar neighbourhood
Galaxy and supernova remnents
wavelength gives very poor resolution, need big or lots of telescopes
Ariceibo telescope - largest size possible for radio telesocpe
Long baseline connects telescopes across the globe
Combines data collection and timekeeping w/ atomic clocks to reference collection
Now have russian radio satellite to add to collection points

Also providwa solar system GPS to track satlelites

1GB a sec from each satellite - LOTS OF DATA!

Previously used tape to collect data
- late 90's several PB a year moved

Move to real time data
- Use Internet to transport data in real time
- Cheaper?

Global network now in operation
- Dedicated lightpaths
- VPNs
- Optimized for transport of data

No hard disk - straight to correlator and processed

Move to 4GB/sec data transfer
- 10 telescopes in operation!
- Yet still shop lots of magnetic data around and want to remove this

Work with SKA in South Africa
- 256 dishes - 90Gb/s per dish

Netherlands station of dipoles
- 240GB/s collection when working!
- 6PB a year growth of archive

120 Tb/s
- 100 days a year
- 130 Eb/y
- 130 Pb/y of processed data!

Question - do we need this data!?
- 99% is noise - can be compressed
- Need all of it to create final images

Question - what do you do with packet loss?

Most of data is noise so not that much of a worry
Use of data is single, not needed to be mined

VLBI = Very Long Baseline Interferometry
(wikipedia: In VLBI a signal from an astronomical radio source, such as a quasar, is collected at multiple radio telescopes on Earth. The distance between the radio telescopes is then calculated using the time difference between the arrivals of the radio signal at different telescopes. This allows observations of an object that are made simultaneously by many radio telescopes to be combined, emulating a telescope with a size equal to the maximum separation between the telescopes.)

the data transport uses UDP but the loss of a few packets is not important as the raw data from the telescopes is essentially white noise. The useful information is obtained by correlating the samples from each telescope.

Question (EGI): how will the storage/compute infrastructure for SKA be procured? how can european e-Infrastructures support LOFAR, SKA etc. i.e. providing services to tackle the future ICT challenges?

For SKA we believe there will be an open tender in 2017 with construction planned to start in 2018. each of the SKA elements is currently doing a Systems Engineering approach from specifications/requirements through concept design and then down select. Some commercial partners are already participating with the element design collaborations.

Question (EGI): what are the computing needs for processing of the raw data? which type of computation is needed? high throughput, high performance, platforms for data analytics on cloud?

Current VLBI uses both FPGA based and software based correlators to proces the raw data from the antenna. SKA is still considering the possible options but FPGA seems attractive.

Round-up of Day 1

Rafa

Genomics becomes distributed and cheap
Hospitals become major generators

Proteomics
- 1EB over Europe?

Imaging
- Community focus on a small section of whole population of producers

Storage vs Publication of data

(DANTE) The network problem is complex in itself. Even a network access upgrade can turn out to be useless, or to not stand up to the expectations, if it's not performed within a well-managed and properly sized global infrastructure.

2.2 Blue sky solutions for big data

If you have any questions or comments please write them below

Genomics federated clouds

data generators not the infrastructure specialists
- what to sequence what to store
- Earlier data reduction stage
  - lowering of bandwidth

Finding the data you want is an infrastructure problem
- Slicing of data needed!
- Can this be something we addressing
- Better management of information
  - New indexing
  - Better metadata
  - Global infrastructure
  - Not just minimum standards
- Not just throwing resources at the problem
- Think as small slices of data not massive datasets
  - Variants not genomes etc
- New computer algorithms to tackle distributed data analysis
- How to compute diverse data sets
  - Unpredictable data set
- ZOOMA curation tool

Takes human curated annotations and applies to new data via computer learning methods

Think smarter not bigger

Exploitation of metadata
- New metadata that relates to infrastructure
- Aid movement and analysis of data
- Scientific + e-infra data

- Semantic metadata
  - Goble + Newhouse paper - REFERNCE

- Not just producers but archivers need to change
  - Can't compute at archive, dont have archive at compute
  - Subsets of data need more tailoring to science questions
    - vice-versa with infrastructures
- RDF Semantic scaling of hardware
  - Specialised equipment for specific problems
  - Q for e-infrastructure and industry
    - How do we drive tech change?

Scale of federation
- Physics - put all data and tools online

Not all producers want to manage data
- Some will
- Empower users to curate/manage their data better
- Federation limited by number of people

Competence in managing reference data from a federation?

Example from Internet - lots of unskilled communities have federated data

Imaging

- 5% of data produced are useful for research
- 95% data destroyed (data not useful/poor quality)
- One center as a reference model
- Each center contributes to the reference center (20%)

- Network capable to manage volumes

Possible solutions - Steven Newhouse

Distributed connected infrastructure
    learning lessons from high-energy physics

Resources in the cloud allow lots of basic research
    Allow for new tools and services to be deployed
    Provide data to assist in locality issues

Build on existing tech rather than reinvent

Key areas - Use case driven

Platform for delivery that can be used!
    How do you deploy services?

What is needed to enable this?
    Where is it
    How to move it
    What to move when it is needed and to the right place

einfra 1
    Closing in September

CB - joining up scientists to e-infra providers

    Avoid diverging
    How do we make sure the vision remains suitable and builds on this discussion

SN - driven by use cases
    Biology has no LHC and Higgs to find
    Means there is a lot of siloing
    Needs commonality to drive Elixir + other ESFRI

Per
    Elixir has a bottom up approach
    Not looking at big architecture

ID needs from members and partners
Lots of national and regional variation

e-Infrastructures need requirements from BMS RIs
Need to be described by user community
Don't develop because you can, develop because you need it

Core development and design
Solve specific problems rather than forcing together solutions

Need to raise awareness amongst basic community scientists of infrastructures

Use of common services
Can platform dev be shared with ESFRI and e-infra?
Federated clouds become very specific
Workflow is community specific
But tech behind can be generic?
Synergies between RIs
EUDAT open calls
Projects could be coordinated between infrastructures
Storage/Compute/Network and Science

Technical and service solutions also needed
Get users to exploit resources
Making services available to a majority of users
Research centers building services for themselves
Research funding is also limited in how it can be spent

Investment into capital
BBSRC capitalising software development
Talk of similar for training
Need to make strong arguments to make the happen

- Need for better organisation of biological data and to speed up current efforts
- Standard in data formats and definitions and length of storage etc to aid industry and service providers in delivering for the community
- Don't look to solve storage/transfer/compute issues as these will be driven to improve - instead focus on describing our current and future needs to push the technology and providers to solve them for the community

- Change the view of how funding works and how it is used by life science ← please expand on this/explain?

Big data checklist for life science infrastructures

Forward look: 5 years from now

Core questions: what is it that BMS RIs need to be thinking about/planning for? Define requirements!

1. Storage: where? what? who accesses? how often? how many simultaneous users? replication? (backup, remote sites?) network requirements for moving/analysing etc.?
2. analysis/compute - where?
3. raw data processing requirements?
4. clouds: commercial? academic?
5. federation? scale of federation? who?
6. curation: who?
7. open data?
8. security requirements?
9. production: where? who? how much? rate of production?

Additional questions:

10. how to define requirements in a useful/understandable way?
11. how to ensure necessary expertise at or translation between data producers/archivists and infrastructure providers?
12. how will information about data be managed and by whom? (both scientific info and e-infrastructure-relevant info)
13. research questions are unpredictable - how much flexibility is needed? (what data will be compared in future?)
14. what technological change to accommodate growing/developing data needs may be needed or desirable? what can we (RIs and e-infrastructure) drive? (e.g. RDF machines)
15. are there ways to rationalise/automate long-term data management/storage (e.g. automatic deletion after “embargo” period)?
16. are there commonalities/synergies between different BMS RIs concerning data that can be exploited? who could lead this effort?

Proposed actions following the meeting

Training

- Teaching users how to efficiently use resources available and improvement of existing resources to make it easier to use them. Lower the threshold
    - Common training proposal across the e-Infra
- Data management training at the point of generation

Support data sharing

- Can life science RIs and e-infrastructures develop a joint proposal for how to facilitate compliance with H2020 Open data pilot? (e.g. support for data deposition etc., see also BMS RI joint paper on data sharing[1])
    - Influencing funders and policy makers on these issues
- Development of tools to aid curation and annotation of data with meta-information
- e-Infrastructure providers to have a federated effort to bridge problem of researchers with poor IT support
    - Work towards the provision of simple tools for use by scientists (e.g. tools around data deposition)
    - UI is very important
- Integration of infrastructure to allow long-term data deposition

Support with sensitive data

- Look at a way to address the need for leveraging EU medical data from multiple sources and in different languages
- Look for support on existing technologies

Develop pilots

- Joint e-infrastructure open call to drive thinking within ESFRI on common issues
- Need to build consensus on use cases and then derive an architecture to iterate new proof of concept studies given the state of e-infrastructures
    - Get feedback from ESFRI on this architecture model and how it fits to data requirements in their community
        - Rafa and Steven to work on this
- E-infra to advance future needs of users by addressing user scenarios
- Help researchers with projects of interest to make the most of the resources available to them
- Uses cases for the short term solving of issue
- Track the science that arises from these resolutions - vertical stories of success in the short term
- Look for the commonalities and rally community to solving these joint issues
    - Technical boards to lead the efforts of communities to address these issues
- E-infrastructures to get together to address issues and look at use cases
    - e-Infrastructure open call on these issues
- Proof of concepts of the capabilities from e-Infrastructure in delivering science services
- Small proof-of-concepts to demonstrate to community that the tech exists and can be deployed to help them
- Well defined use cases - must be representative of problems that need to be solved

---

[1] ELIXIR *et al* (2014). Principles of data management and sharing at European Research Infrastructures. ZENODO. 10.5281/zenodo.8304

- Education of users + scientists on e-Infrastructures and what they can provide
- GEANT - need single/several use cases from science community as a model for other efforts to solve data deluge
    - ESFRI (BMS RIs) to work on these for e-Infra
- Use case DANTE-Euro-BioImaging (Jason S.)

Communication/meetings

- BMB to facilitate similar meetings between national stakeholders in ESFRI RIs/e-infrastructures etc or to raise the profile of these meetings to the community to allow greater understanding of the use cases by science and EC
- Liaising between scientists and IT services to make use of what is there - nurture experts
- Interest groups created that focus on needs
    - These could lead their own meetings so better focus on topic at hand
- Regular meetings of the main group, e.g. bi-yearly meeting?
    - see e.g. radio physics - working group for mutual exchange between IT and science community; Series of meetings with regular updates to the community
- General meetings to continue the elements that have come out of this workshop?
- Next meeting around RDA meeting (end of Sep, Amsterdam)
    - set up RDA workshops?
- E-infra meetings of interest groups
    - Use case to pilot case can be driven off these meetings
- Proposed single sign-on workshop in September
    - Not the raison-d'etre of this group ← who decides?
    - AAI might be a better focus - might not be worthy of an entire workshop
    - Invite in 'off the shelf'
    - Both of these are solved - more an issue of dissemination to user body; Possibility of training events for these
    - Need to be clear what is solved: technical approach with comparatively low-level security: yes, adoption: no, single sign-on to sensitive data sources: no ← latter most relevant for BMS RIs
- The life science disciplines have similar needs concerning storage, moving data, access etc.
    - Decide who needs to know about and action IT infrastructure issues
        - Is it ESFRI or not? NO - the drive would have to come from the RIs; funding to be determined. (Unless "ESFRI" here does not refer to the body, but the research infrastructures - need to be clear on terminology)
    - Formalised attempt at agreement on common ground and differences in life science data to better drive solutions to tackle the big problems
    - ID the impacts that these issues have on the science pipeline and where they occur

- Need to keep focus on the science questions we need to answer! Clearly state the open questions and gaps

Other

- It would be helpful to have a brief head-to-head comparison between the e-Infrastructures: what problems should typically be addressed by the one or the other? From the outside there appears to be overlap.
- Evaluate writing of paper on this topic
  - Use notes and board cases on problems/solutions from the meeting
  - Need a detailed document behind a paper to fully cover the knowledge and information to the community
  - Document needs to be a 'living' one that will change as the science and tech evolves

Timeline for next steps

Draft vision from community

December forum to discuss the vision

Resource list

Please use the space below to provide links to resources that might be useful for future reference.

Google Map of NGS machines worldwide - http://omicsmaps.com/

XKCD What If of FedExNet - http://what-if.xkcd.com/31/

Programme

Time	Session	Speaker
Day 1		
12:00	Arrival and lunch	
13:00	Introductions	Tom Hancocks Stephanie Suhr
13:10	Challenges of big data & aims of the workshop	Rafael Jimenez
13:30	Data challenges of different science communities	
	Genomics	Pieter Neerincx
	Proteomics	Henning Hermjakob
	Imaging	Jason Swedlow
	Metabolomics	Natalie Stanford
	Clinical data	Jan-Willem Boiten
14:00	Data fluidity	Guy Cochrane
14:20	Flash presentations: e-Infrastructures	
	EGI	Tiziana Ferrari
	EUDAT	Per Öster
	GÉANT	Richard Hughes-Jones
	PRACE	Sergio Bernardi
	CERN/LHC	Alberto di Meglio
15:00	Round-up of challenges	Rafael Jimenez
15:10	Open discussion/Questions and Answers	
15:30	**Break**	
15:45	Science community use case	Group session
16:15	Report back from group session	
16:45	Science community use case	Group session
17:30	Report back from group session and discussion	

18:30	End of day	
19:00	Dinner at the Red Lion, Hinxton	

Time	**Session**	**Speaker**
Day 2		
09:00	Solutions for big data in other science communities	
	Earth satellite data	Wolfgang Lengert
	Radio astronomy data	Arpad Szomoru
09:30	Blue sky solutions for big data	
		Group session
10:30	Report back from group session and discussion	
11:00	Break	
11:30	Practicalities and actions to implement solutions	
		Group session
12:30	Report back from group session and discussion	
13:00	Lunch	
14:00	Closing discussion, roundup of challenges and solutions	
		Rafael Jimenez
15:00	End of workshop	