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

A BioMedBridges knowledge exchange workshop hosted by ELIXIR, *15-16 May 2014, Hinxton, UK*

Report authors: Stephanie Suhr, Guy Cochrane, Nathalie Stanford, Jan-Willem Boiten, Jason Swedlow, Chris Morris, Rafael Jimenez, Pieter Neerincx

SUMMARY

This BioMedBridges knowledge exchange workshop[1], held on 15-16 May in Hinxton and hosted by ELIXIR, provided a forum for the biomedical sciences research infrastructures (BMS RIs) to discuss their possible future service requirements with e-infrastructure providers including GÉANT, PRACE, EGI, and CERN Openlab. The workshop addressed biological 'big data' and the challenges to be faced by the life sciences five years from now and further into the future, focussing on the needs for storage, transfer and computation of biological data.

[1] Slide presentations from the workshop are available at http://www.biomedbridges.eu/trainings/knowledge-exchange-workshop-preparing-data-deluge

Data growth will change how the life sciences work today. Although different communities (currently) use different models, there may be common approaches and solutions that must be identified. At the same time, existing technology must be used more efficiently and the life science community has to evaluate the practicality of storing everything. Data protection/security and issues around sensitive data (e.g. patient data) make life science big data more challenging.

There are an opportunities both for e-infrastructures to better understand data-related challenges of the life sciences (BMS RIs), and for the BMS RI to get better at defining their requirements. Practically, at least part of the initial engagement with e-infrastructures should be local/bilateral between BMS RIs and e.g. the corresponding national nodes of individual e-infrastructures; a more generic overview can then be achieved once there are working examples or concrete use cases. Overall, there is a need for capabilities development in the life sciences concerning how to deal with scientific data and IT services that must be addressed.

Introduction

The data deluge in the life sciences is vast and growing increasingly fast, resulting in potential bottlenecks including data production growing much faster than storage, the cost and availability of data production technologies (e.g. whole genome sequencing) declining faster than the cost of available storage technology, and the fact that it can take longer to transfer data than to produce it. It is obvious that there is a need for the life science research infrastructures to work together in addressing these challenges, and that there would potentially be great benefits in drawing on expertise by the existing e-infrastructures to assist in this.

However, it is not trivial to determine specific opportunities and how they may be addressed as the actors involved both on the life science and e-infrastructure sides have so far not spoken the same language in technical terms. In addition, there is some great variation concerning data and possible approaches even among different life science disciplines, such as for example concerning genomic data and imaging data. The “usual suspects” may not be the most urgent areas to focus on - for example, DNA sequencers can produce ~100 GB of data in 24 hours, while the most sophisticated microscopes and mass spectrometers could produce up to 4 TB in the same period.

The outdated, casual approach to data sharing “the data is on my disk and available to anyone who requests it” has gradually been replaced by submission to specialist data repositories. However, the data deluge produces new challenges which threatens the data sharing process and complicates the access, integration and analysis of data. For example, as genomics becomes cheaper and is more routinely used in clinical research, patient care and treatment, hospitals become major data generators, leading to immensely disperse distribution of large, valuable data sets and resulting in questions of where to store, process, and how to transfer the data. Potential solutions relate to storage (data compression, selective storage), networking (faster protocols, partitioning, network upgrades) and computation (clouds, locating data close to computation.

Data challenges of different life science communities

The life sciences are diverse and include many domains with many different data types, data management procedures and data sharing cultures and requirements. To make the workshop as productive as possible, a small number of representative domains were invited that have already started to experience difficulties in dealing with the data deluge.

The representatives of these domains were asked to provide information on:

1. Domain and data types: brief presentation of the domain and the data that is generated
2. Data production
    a. How much experimental data can be produced by average technology vs. the most sophisticated technology?
    b. Is it possible to reproduce data from reprocessing the same biological samples?
    c. What is the average cost of producing the data?
    d. What is the degree of distribution of data production?
3. Data processing
    a. Sometimes raw data coming from instruments undergo different types of transformation or post-processing (e.g. format transformation, compression, integration, etc. What are the data processing requirements and the time and computational resources required at this stage?
4. Data submission
    a. What is the submission procedure to public repositories - what data is submitted (raw data, processed data, metadata), what is the data volume, is it an automated process?
5. Data repositories
    a. What is the growth of data in public repositories and its trend over the last few years? How much of this data is raw data?
    b. How many public repositories collect the data in your domain?

c. Is there any kind of existing collaboration to exchange information, replicate the data or work on a federated model with distribution of responsibilities?
d. Are there any mechanisms encouraging data submissions (e.g. journal publisher requirements or data management plans from funding agencies)?
e. How much of the data produced by scientist in the discipline ends up in public repositories?

Genomics

Big data offers unique possibilities for personalised medicine. As genomics is becoming more routine in clinical environments, one of the main challenges with genomic data is data protection. There is an argument that whole genome data itself is identifiable, given sufficient context; however, genomic data can be (and is) shared given adequate precautions (informed consent, security measures, etc.). Another solution to this is not to share data about individuals, but only about groups (aggregated information, e.g. “all male patients over 50 years of age”).

The (lack of) quality of clinical annotations (i.e. careful and standardized description of the phenotype) is one of the major concerns with current genomics datasets, which makes them less useful for clinical purposes than they could be. This is even the case for some of the most well-known genomics data collections currently available .Further challenges revolve around quality control and management of large data sets. While summary statistics are insufficient, manual inspection of datasets is not feasible.

The type and size of data sets varies, ranging from 10-100 MB per file of variant calls via an entire genome of raw FASTQ reads of about 100 GB to many whole genomes (several TB). As research and clinical care go hand in hand with new technologies, data must be kept close to patients and secure.

In clinical research projects, data often has to be shared between many different locations. In the case of large genomics datasets, connectivity issues then

become very important. There is no real-time data processing and analysis requirement; however, in general, turn-around times of genomic analyses and results are very important for patients; these times can sometimes still be long and must be shortened in future (current time from whole genome sequencing to diagnosis <7 days, aim is reduction to <2 days before end of 2014). Securing appropriate performance demands local data storage and access; however, this may make integration of data with other resources more difficult.

Proteomics

PRIDE hosts mass spectrometry-based protein expression data and has two main aims: to serve as a repository (provide data supporting proteomics publications) and as a source of proteomics data for other data resources.

Proteomics is marked by a huge diversity of experimental and data processing approaches. It is not possible to develop and maintain parsers and proper database representation for all approaches, but only for the most important ones. Consequently, until 2012, only those submissions to PRIDE were accepted that could be fully represented. Since then, “partial submissions” have been allowed - these still require raw data, processed data, and metadata, but are handled as structured collection of files, only metadata is stored in a structured manner. This has improved the service to the community, but also bears the risk of users doing partial submissions because they are easier/less work intensive.

In ProteomeXchange, the number of submissions per year has been steadily rising, with 102 in 2012, 527 in 2013 and 192 only in the first quarter of 2014 and datasets covering a total of 215 different species. The total data volume of PRIDE is >40 TB and looks to be growing exponentially since 2012, with the largest single submission 4 TB and the total number of files >120,000.

A specific challenge of proteomics data is the transfer of large files (upload/download). Currently this is supported using commercial protocols (Aspera), which are costly.

Imaging

Imaging data might present the biggest data challenge in the life sciences in terms of volume: while imaging technologies are getting better and much more widely available also in clinical settings, image compression algorithms are computationally expensive and hard to standardise across a diverse range of imaging modalities and applications in clinical, industrial and academic research institution. In many cases, lossy compression schemes introduce artifacts that may affect the outcome of downstream analyses, so the community often avoids compression schemes in favor of simply accessing more storage.

As an example, digital pathology at a single hospital site can currently generate up to 2 TB per day, and at only €250k per machine the number of da-a producing sites is growing rapidly. Similar scales are achieved in academic research labs. As a result, surveys of users in academic and clinical laboratories repeatedly cite data management, processing and analysis as a limiting factor in their productivity and efficiency.

The imaging community has a need for an annotated repository of image processing and analysis tools as well as an open and accessible image data repository, providing open access to standardized, annotated data or reference image data (~1000 TB in yrs 1-2). Currently, data transfer methods are determined by the size of the dataset and may consist of web-based transfer or shipment of external hard drives. To deliver its maximal value, image data must be integrated with other relevant data such as molecular resources (e.g. GWAS phenotypic screens) or structural resources (e.g. super-resolution, correlative).

In the context of Euro-BioImaging, the volume of user-generated image data at each producing node is anticipated to reach up to 20 TB per year on average (for an average of 30 users per year at each node), with 80% of users requiring <200 GB and 20% need 0.2--20 TB storage. The plan is for initial data storage at the Euro-BioImaging data production nodes for quality control and initial processing; afterwards, the data belongs to the user (i.e. the user must arrange storage).

Compute capabilities are needed for data mining and the development of new tools. A range of storage and compute solutions will be required to match the diverse needs of the community of imaging scientists. In most cases, dedicated storage and compute capacity are directly linked to each image data acquisition resource. The first generation of linked, annotated, public image data resources are now available and have demonstrated technical feasibility and utility. In the future, both compute and storage can be implemented in academically-owned, cloud-based solutions that can host algorithms/tools (VM) and are linked to repository/benchmark data, bypassing the need for data download by the user. In a pilot study, the Helix Nebula project2 is testing academically-operated science cloud computing for Europe, with EMBL providing the first life science use case.

A solution to the community image data repository in the longer-term may be a centralized catalogue instead of a single (big as you like) system. There are issues around coherence of the data between the various sources/sites, but other fields may provide models for solutions (e.g. LHC ATLAS distributed data systems).

Metabolomics

Metabolomics involves the profiling and quantification of metabolites from mass spectrometry (and NMR) data. The amount of raw data generated per experiment using mass spectrometry is ca. 15 MB per single sample/15 GB per batch of 1000 samples and the derived data is 10 kb per batch. Data processing can take up to 17 days for 1000 samples, depending on available compute.

There is a large variety of data (sources, forms, structures). To ensure data quality, standard operating procedures are used throughout the analysis pipeline (for example, quality assurance/quality control is applied to check instrument drift). Metadata such as sample methods, ChEBI ID and Inchi string are added so data can be reused.

[2] www.helix-nebula.org

As technology advances, much of the available/stored data needs to be replaced (data improves short-term: 2 year cycle). Raw data is not dismissed as it can be re-analysed in many different ways for many different purposes. As instrumentation becomes more powerful in detection, more metabolites can be identified per individual study. There are currently ca. 20 listed resources for metabolomics data.

Metabolomics often uses NMR as well as mass spectrometry - these data have different formats, rates of generation and sizes to the proteomics data, but may have to be analysed together with the mass spectrometry data as part of the same study. The volume of raw data depends very much on the assays run per sample, ranging from a few GB per assay to several hundred GB for complex assays. This can result in annual outputs of several PBs per year for a large facility. There are extensive pre-processing steps, often requiring cross-referencing of datasets, and local quality control standards.

Estimates of the numbers of sites generating metabolomics data across Europe five years from now are ca. 5 large facilities and about 100 other data producing sites (but note that numbers are likely inaccurate). A recent survey conducted as part of the ISBE preparatory phase identified that about a third of systems biologists expected to use some type of metabolomics in their research in the future. Expectations for data volumes five years from now are 5 PB raw data, processed at GB level.

In a number of countries, very large groups are starting to set up national resources that will have 20-30 mass spectrometers as well as NMR spectroscopy (the current number of machines is smaller, with either mass spectrometry or NMR being available, not both).

In the metabolomics community, researchers make use of lots of resources to re-analyse data. Challenges include restrictions on movement of data (legal/ethical problems of data sharing), quantification against datasets from different sources/locations, varying standards from different technology platforms and a great deal of variation in software and file formats. At present, a lot of data ends up in small, experiment-specific data repositories. Although there is a push to use

central repositories, in reality there is no one size fits all solution - there will still be requirements also e.g. for local storage, processing etc.

Clinical data

For clinical data - sensitive patient data - security and trust in archiving is key. From the IT side, solutions are needed to keep data and IP safe in order to build trust. User interfaces and usability of tools are also very important as the tools will (or should) be used by clinicians, who are often not data- or even bioinformatics experts. Compute is needed for image processing and processing pipelines for DNA/RNA sequence data. Many issues in this area are already covered by the genomics community.

A large amount of data produced for clinical care is not research data per se; there is still a large divide between clinical care and research. The overwhelming culture in the clinical context is risk averse, and there is a reluctance to data sharing - clinicians often prefer to keep patient data on the premises (in the hospital) and like to remain in control of the data. Big driving factors here are politics and fear of litigation; however, the biggest difficulty is that clinical/patient records, reports and demographics are recorded largely in natural language instead of controlled vocabularies, which would be fairly easily translated to English. While clinical care focuses on the careful description of the individual patient, clinical research requires registration of the data in a controlled manner in order to allow comparisons across the patient cohort. In addition, physicians will not record more data than required (time is money for them), while research usually requires a much richer data set. To address this, each individual hospital must work towards data interoperability concerning their patient records and data annotation.

Stored data includes both raw and processed data. Monitoring or study of disease progression or development involves long-term collection of data. It can be difficult to compare patients with similar conditions (patient stratification - difficulty e.g. inconsistent data annotation/terminology used in patient records). Five years from now the volume of data per patient is estimated to be about 10 to 100 GB.

Important developments include the interest of insurance companies in healthcare data to 1) benchmark healthcare providers in terms of quality of treatment and 2) analyze outcome data in order to determine the effectiveness of treatments, as well as the increase of personal monitoring devices and direct-to-consumer analytical services, which result both in patients becoming more involved in their own health care as well as providing a possible additional, very rich and informative data source.

Data fluidity

Background

The cost of sequencing has fallen dramatically, leading to the major bottleneck in life science research today being data analysis. Users of Europe's biological databases range from clinical specialists via environmental researchers to computer scientists. Storage and processing of large volumes of data have become a challenge.

As an example, EBI databases experience exponential growth, with the volume of data doubling at times over the last few years as frequently as every 9 months. The current volume of EBI's data resources is 5-10 PB, which is replicated at write time - this works well and fast as there is substantial network connectivity between the EBI data centres. Total storage with backups is ca. 40 PB. The EMBL-EBI website is visited by approximately 11,000 unique IP addresses a day (note that this number could represent many more users as IP addresses sometimes represent entire organisations). All bioinformatics resources in Europe together currently have upwards of 60 million hits a month. Data that is deposited in the EBI databases is generated worldwide by small instruments (in comparison to those of other scientific communities) producing data in an manner that lacks any single central organisation.

It is important to note that both the cost and the relative (qualitative) value of samples (data) depends on the entire process of data generation, including its origin, sampling, manipulation/experiment, etc. While the cost of sequencing itself

has become very low, other related costs can be (very) high or even immeasurable, depending for example on the conditions under which a sample is taken or whether the sample is unique. In practice, data can often not be regenerated: in the case of patient data for example a sample cannot be regenerated or re-taken when the disease to be studied has progressed and the initial sample has been expended. Similarly, there will be huge differences in the value of research data gained during oceanographic or polar research campaigns (expensive, difficult or even impossible to repeat) and laboratory-based experiments that can more easily be repeated. Whether a sample is reproducible depends on the scientific context, and what is stored, how and for how long will ultimately have to be driven by the respective scientific disciplines, who are the experts on the inherent value of the data in question.

Life science data storage solutions

Data compression. This requires models to ensure that lossiness is minimised while achieving the desired economies. Compression itself can be computationally expensive. CRAM is a compression method used for sequence data that uses well established algorithms (eg. RLE, Golomb-Rice, etc.) and structures sequence data appropriately to leverage these. In lossless mode, some compresion is achieved while in lossy modes, very aggressive compression is possible. CRAM cannot be used directly for data other than sequence data as it uses sequence data-specific characteristics. However, the reference and variation from reference basis is generic, leveraging image and video compression approaches.

While working practices are suggested in which data are held initially in uncompressed forms and then, at a later stage, reduced under some lossy compression model, limited overall cost savings are made with this approach: the cost of maintaining the dataset in question will have decreased significantly (exponential growth of disk capacity per unit cost): in volume terms, legacy data are not nearly as important as future data.

Data partitioning. Freeing up network and IO bandwidth by partitioning will create more space for submissions, follow-up and downstream analysis of large data sets (depending how well joined up the services are). It is important to ensure that data is stored and dispatched in a way that best addresses the user needs (the biological context must be maintained): in genomics, there is a clear reference genome model, while in environmental sequencing the indices may be functions (around gene/pathway) or taxonomic (around taxa and clades).

Cloud computing. To address the issue of available compute for very large volumes of data, or where downloading data is not feasible due to volume, EBI and ELIXIR are piloting cloud services ('Embassy' project) co-locating compute with data storage: data can be mounted directly onto these instances for direct compute. An EGA (controlled access data) service from CRG Barcelona was also launched in 2014, in part to bring data closer to compute available in e.g. the Barcelona Supercomputer Centre.

In the case of patient data, where some legislations will prevent any non-domestic data export, cloud computing may not be possible. However, there are discussions in the Global Alliance for Genomics and Health[3] initiative, for example, on the possibility of making some level of summary/aggregate data available. In addition, if legislations allow, a cloud "Embassy" - where a nation has ownership, control and privacy despite the cloud technically being on foreign territory - is technically available. However, this issue could be influenced also by soft factors, such as "public opinion/perception" (see e.g. reaction in Germany to Google Streetview[4]).

[3] http://genomicsandhealth.org/

[4] http://bits.blogs.nytimes.com/2013/04/23/germanys-complicated-relationship-with-google-street-view/?_r=0

e-Infrastructures

The e-infrastructures were asked to highlight their respective capacity and services to deal with the data growth in the life sciences, specifically with respect to:

1. Storage, transfer and computation of data
2. Available funding or funding models the e-infrastructures anticipate using to provide solutions to research infrastructures
3. Existing resources the e-infrastructures can offer, their current usage, the limitations and plans to deal with the data deluge.

In looking at these services, the following stakeholders involved in the process of sharing data need to be taken into account:

- the scientists/institutions producing the data
- the public repositories collecting and integrating the data
- the scientist downloading and analysing the data.

Solutions for big data

Two presentations provided examples for complex data management requirements in other disciplines for comparison.

Earth satellite data

Earth Observation data is not really “big data”, but many small sets of satellite data and involving reception, processing and dissemination to research centres. Monitoring of Earth focuses on geo hazards (earthquakes and volcanos) and includes radar, sea-surface temperature, ocean colour, gravity, electromagnetic and other types of data.

Data volumes vary depending on the type of data in question, ranging from a couple of TB via about 80 TB for sea temperature and sea level data and PBs for oceanographic data.

Space agencies have an open data policy; making data available to universities and for research requires infrastructure. There is a move towards a generic infrastructure for a multi-tenant provider (= big pool for data supply/use). Google is keen to get involved in this, but the Helix Nebula is a way to retain control over both the infrastructure and the data. In this specific case, the technology actually drives the science.

Radio astronomy data

The VLBI Network combines radio telescopes as one instrument 70 days a year. JIVE oversees this collaboration and brings data together. The telescopes capture cm wavelength of the electromagnetic spectrum to explore the stellar neighbourhood of Earth and describe galaxy and supernova remnants. Since the wavelength gives very poor resolution either very big or a lot of telescopes are needed. The Arecibo telescope has the largest possible size for a radio telescope.

A long baseline connects telescopes across the globe and combines data collection and timekeeping w/ atomic clock data to reference the collection. In addition, solar system GPS is provided to track satellites. The data volume is 1GB per second from each satellite.

The global network now in operation has dedicated lightpaths and VPNs and is optimized for transport of data. Data goes straight to the correlator and is processed. The plan is to move to 4GB/sec data transfer. In comparison, SKA in South Africa (256 dishes) captures 90GB data per second per dish, and the Netherlands station of dipoles 240GB per second during operation. The total amount of data is 130 PB per year of processed data. Although 99% of this data is noise and can be compressed, all of it is needed to create final images.

Conclusions

Specifics of life science data

- Compared with physics, biological data is much greater in variety in terms of the types of data. Production, storage and consumption. The question of storage of very large volumes of data is independent of the discipline; however, storage of raw data may be more important in the life sciences than in physics since in the latter data is usually processed already during capture.
- Users access data in a different way: IO is a key factor in life sciences - profile of access to databases shows that ca. 40% of data is accessed soon after submission and there is frequent repeat access
- Comparisons of data resources with other data resources are frequent, including big with big
- Life science disciplines such as genomics have a rich approach to accessing annotated and integrated data
- Data protection/security and issues around sensitive data (e.g. patient data) can make life science big data more challenging - for example, it can be difficult to separate big data from computation due to restrictions to export of sensitive data
- Distributed connected infrastructure: there is an opportunity to learn lessons from high-energy physics (but biology has no LHC and Higgs to find, which currently results in siloing)
- The life science disciplines have similar needs concerning storage, moving data, access etc.
    - Need agreement on commonalities and differences in life science data to be able to drive solutions for big problems together
    - Need to identify the impacts that these issues have on the science pipeline and where they occur
    - Need to focus on the science questions that need to be answered - clearly state the open questions and (technology) gaps

Technology development

- Technology may not be a problem: new technology solutions might become available in time (e.g. disk companies are coming out with new types of disk today that can store many times the current maximum volume).
- Don't look to solve storage/transfer/compute issues as these will be driven to improve - instead focus on identifying and describing current and future needs to push the technology and providers to solve them for the community
- The community needs to request better e-infrastructure: involve the user community in identifying needs
- Interactions are needed at every level; data generators are often not infrastructure specialists
- Trust of researchers using infrastructures may be an issues - do users trust e-infrastructures to provide or build the right services for them? If the trust is not there, communities might prefer building their own infrastructure instead of buying a service/using existing infrastructure: this may not be cost effective
- If every data producer must store (at least some of their data) locally, it is probably more a cost problem, not technological. Having centralized storage might help to bring down the cost in addition to allowing easier data management.

Factors influencing data availability

The workshop participants determined the following factors that can influence data availability:

- **scientific** (e.g. data reproducibility, uniqueness, value of processed and/or raw data)
- **financial** (cost of data storage, transfer, reproduction)
- **technical** (storage, network, computation…)

- **political** (drivers e.g. from funding bodies/large organisations/national interests)
- **social** (data sharing mentality of the community in question, how they do science - e.g. standards, best practices)
- **legal/ethical/formal** (requirements/constraints for data storage/transfer/access – e.g. need to store patient data in country of origin, requirements from journal publishers, data management plans, etc.)

"Big data checklist" for life science infrastructures

Participants were asked to take a forward look five years into the future and anticipate the information that would need to be available in order for effective support to be provided to the life science research infrastructures. The following questions emerged concerning what BMS RIs need to address to define their requirements and be able to start a productive dialogue with e-infrastructure providers.

Core questions

1. **Data storage, volume and access**: Where will data be stored, and what data will be stored? Who will access stored data and how often? How many simultaneous users? Will stored data be replicated (backup, remote sites)? Are there ways to rationalise/automate long-term data management/storage (e.g. automatic deletion after "embargo" period)?
2. **Networking requirements**: what are the network requirements for moving data - from production sites, for storage, analysis etc.?
3. **Data analysis/compute**: where will this be located?
4. **Raw data**: what are the RI's raw data processing requirements?
5. **Cloud solutions**: If desired, what would be most suitable: commercial? academic?
6. federation? scale of federation? who?
7. **Data curation**: who will curate the data? Are the necessary experts available?

8. **Open data**: will the data be open to the wider research community, or accessible for the relevant scientific community? Under what conditions? What will the level of openness be? Are there restrictions?
9. **Data protection and security**: what are the requirements?
10. **Data production**: where will data be produced? By whom? How much? What will the rate of production be?

Additional questions

1. How can requirements be defined in a useful/understandable way? (What are the key questions that can be addressed with possible solutions?)
2. How can the necessary expertise at or translation between data producers/archivists and infrastructure providers be ensured?
3. How will information about data be managed and by whom (both scientific information and e-infrastructure-relevant information)?
4. Research questions are unpredictable - how much flexibility is needed? (What data will be compared or analysed in future?)
5. What technological change is necessary or desirable to accommodate growing/developing data needs? Can this be driven by RIs and e-infrastructures (e.g. RDF machines)
6. Are there commonalities/synergies between different BMS RIs concerning data that can be exploited?
7. Can life science RIs and e-infrastructures jointly influence funders and policy makers on these issues?

Proposed actions following the meeting

Training

- Teach users how to efficiently use resources available (utilise a “train-the-trainer” approach)
- Provide data management training at the point of generation
- Educate of users and scientists on e-Infrastructures and what they can provide

Support data sharing

- Improve existing resources to make them easier to use - lower the threshold
- Develop tools to aid curation and annotation of data with meta-information
- e-Infrastructure providers to develop a federated approach to bridge the problem of researchers with poor local 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

- Address the necessary requirements for leveraging EU medical data resulting from records being spread over multiple sources, in natural language and with incomplete information, and in different national languages
- Look for support on existing technologies (data security, support with legal and ethical requirements)

Develop pilots

- 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
- Well defined use cases - must be representative of problems that need to be and can be solved.
    - These could consist of helping researchers with projects of interest to make the most of the resources available to them: small proof-of-concepts to demonstrate to community that the technology exists and can be deployed to help them - demonstrate capabilities of e-Infrastructure in delivering support to science
    - Track the science that arises from these solutions: vertical stories of success in the short term

- Look for the commonalities and rally community to solve common issues: technical boards of BMS RIs should lead the efforts of their respective communities

Communication/meetings

- 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. This will serve the goal of capability building, by creating a small community of practice that is in touch with developments both in life sciences and in computing.

ANNEX 1: Infrastructures represented at the meeting

Biomedical Science Research Infrastructures

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

e-Infrastructures

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

ANNEX 2: Workshop participants

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