Realising the full potential of research data: common challenges in data management, sharing and integration across scientific disciplines

---

Version: 3

Date: 20 December 2013

Authors: L. Field (CERN), S. Suhr (EMBL-EBI), J. Ison (EMBL-EBI), W. Los (UVA), P. Wittenburg (MPI), D. Broeder (MPI), A. Hardisty (Cardiff University), S. Repo (ELIXIR), A. Jenkinson (EMBL-EBI)

Summary

Established and emerging European research infrastructures are holding, or will be holding in the near future, immense quantities of data. The intrinsic power of data does not only come from storing and managing these data, but from making the data available and accessible to a wider audience, across national borders, scientific communities and disciplines, and by integrating datasets so that more complex scientific questions can be solved.

Such endeavors have challenges, many of which are shared between different scientific communities. To exchange existing expertise and address obstacles, the BioMedBridges, CRISP, DASISH and ENVRI projects - covering the biomedical sciences, physics, social science and humanities, and environmental sciences - have come together to identify cross-cutting topics, discuss current approaches and develop recommendations for future actions requireed to solve them.

The ESFRI Cluster Projects are funded by the European Commission within Research Infrastructures of the FP7 Capacities Specific Programme, grant agreement numbers 284209 (BioMedBridges), 283745 (CRISP), 283846 (DASISH), 283465(ENVRI).

Introduction

The quantity of data held by established and emerging research infrastructures in Europe is immense, and with the emergence of new technologies, such as high-throughput genome sequencing and X-ray free-electron lasers, are growing exponentially. In parallel, the awareness of big data and the absolute importance of data management, processing, analysis and sharing, has increased dramatically over recent years. While scientific disciplines have previously addressed data-related issues themselves and mostly within their own communities, the current and future challenges - which may be technological, sociological and/or economic - have become too large for such a silo-centric approach. In addition, while there are discipline-specific topics with respect to data, it is becoming increasingly clear that there are a large number of shared problems, amongst differing disciplines.

The BioMedBridges, CRISP, DASISH and ENVRI projects have come together in an effort to identify these shared challenges. The projects represent "clusters" of research infrastructures in different disciplines - biomedical sciences, physics, social sciences and humanities (SSH), and environmental sciences - on the European Strategy Forum for Research Infrastructures (ESFRI[1]) roadmap, and thus span a tremendously wide range of scientific communities and cultures. Given this diversity, identifying the commonalities is anything but trivial. In this context, the current publication should be seen as a first working paper that is prepared in an effort to support further discussions.

The four ESFRI Cluster Initiatives

BioMedBridges

BioMedBridges brings together BBMRI (biobanks), EATRIS (translational research), ECRIN (clinical trials), ELIXIR (bioinformatics and life science data), Infrafrontier (mouse disease models), ERINHA (contagious diseases), EU-OPENSCREEN (cheminformatics and chemical screening platforms), EMBRC (marine model organisms, analysis platforms and metagenomics), Euro-BioImaging (biological and medical imaging) and INSTRUCT (structural biology).

The combination of the significant data resources in the biological and biomedical sciences will help answer complex and important scientific questions. However, there are substantial challenges in accessing and sharing data resources across the domains. To address the data sharing issue, the BioMedBridges consortium aims to define, implement and deliver, data interoperability across their domains. Each of these RIs has specialised computational and data resources. Well-defined use cases, the implementation of which depends on the exchange of information across domains and between multiple BMS RIs, drive the development of computational 'data and service' bridges. The use cases include, for example, the identification of treatment options for cancer patients by linking drug screen with genomic data, or translating data between mouse model organisms and human clinical information for diabetes and obesity. A central objective is to implement interoperable standards and ontologies across the different data resources and services to allow correlative

[1] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri

analysis. Public data will be made freely and widely accessible through these standard interoperable services. Where sensitive data is shared, such as medical information or data protected by Intellectual Property, standards for secure and restricted access are identified and implemented.

CRISP

CRISP includes ESRF, EuroFEL, and European XFEL (photon sciences); ESS, ILL (neutron science); FAIR (antiproton and ion research); ILC-HiGrade, SLHC, GANIL-SPIRAL2 (particle physics); ELI (photon science and nuclear physics), SKA (astrophysics).

The Cluster of Research Infrastructures for Synergies in Physics (CRISP) project brings together eleven Research Infrastructures (RIs) within the physics domain. The objective is to build collaborations that create long-term synergies that will enhance efficiency and attractiveness of those RIs. The CRISP project focuses on four R&D topics that are of utmost importance for the participating RIs: Accelerators, Instruments & Experiments, Detectors & Data Acquisition, and Information Technology (IT) & Data Management. In the area of Data Management, new initiatives and approaches are required to cope with the ever-increasing flow of scientific data produced by the next generation of detectors. A joint effort to establish the base elements of adequate platforms for the processing, storage and access to data is required.

ENVRI

ENVRI includes LifeWatch (biodiversity and ecosystem observations), EPOS (earthquakes and volcanoes observations), ICOS(greenhouse monitoring), EISCAT 3D (space and upper atmospheric physics), EMSO (deep seas observations), and EuroArgo (open seas observations). ENVRI also interacts with IAGOS (aircraft for global observations) and SIOS (Svalbard arctic Earth observations).

The central goal of the ENVRI project is to draw up guidelines for the common needs of the environmental ESFRI projects and to implement common solutions, with a special focus on issues such as architectures, metadata frameworks, data discovery in scattered repositories, visualization and data curation. These solutions will empower the users of the collaborating environmental research infrastructures and enable multidisciplinary scientists to access, study and correlate data from multiple domains for "system level" research. The collaborative effort will ensure that each infrastructure can fully benefit from the integrated new ICT capabilities, beyond the project duration, by adopting the ENVRI solutions as part of their ESFRI implementation plans. In addition, the result will strengthen the European contributions to GEOSS (the Global Earth Observation System of Systems). All nine Social Benefit Areas[2] identified and addressed by GEO-GEOSS[3], will take advantage of such approaches.

DASISH

DASISH consists of CLARIN (linguistics), DARIAH (arts and humanities), CESSDA (social sciences), SHARE (research on aging societies) and ESS (European social sciences survey).

[2] http://tinyurl.com/oxs4z95

[3] http://www.earthobservations.org/geoss.shtml

DASISH brings together all five ESFRI research infrastructure initiatives in the social sciences and humanities (SSH). The goal is to determine areas of possible synergies in the infrastructure development and to work on specific concrete joint activities, to facilitate cross-fertilization, harmonize approaches and knowledge where possible. DASISH has identified four major areas of activity namely data quality, data archiving, data access and legal and ethical aspects. With respect to data quality, the big challenge is to find methods that allow better integration of data, in a cross-disciplinary and cross-border setting. Data access covers a whole bunch of different activities, such as establishing a joint tools and knowledge registry to create a common marketplace, establishing a joint metadata domain on data, extending the knowledge about AAI solutions and creating a start-up federation, studying methods and advancing tools for Web-based annotation and studying workflow systems and common requirements. With regards to archiving, DASISH wants to identify operational and trusted deposit services and work on requirements for policy rules.

Inventory of common topics

Initially, sixteen topics of common interest were identified (Table 1). Interestingly, almost all of them apply to all four disciplines: only two apply to all four except physics and one to all except physics and biomedical sciences. As may be expected, the former two (that do not apply to physics) are related to semantics and semantic interoperability and the latter topic (that only applies to SSH and environmental sciences) to dynamic data management. A proposed user community body, reference models, education and training are seen as supporting activities. The definition of these sixteen topics provides an excellent basis for further discussions.

Data identity

It is recognized across scientific disciplines[4] that there is a need for researchers to publish their data and for other researchers to access and cite that data, in a standardised way. Data citations enable the attribution of credit for those who created the data, which itself can be a mechanism to encourage data sharing, while establishing data provenance. Standards and practices to enable data citation are necessary to promote sharing and reuse of research data across disciplines, to ensure the full potential of the data can be achieved.

Persistent Identifiers (PIDs) have been used for many years to identify publications, for example the ISBN and DOI systems. Similarly, all data objects created as a result of the scientific process should be registered and assigned a PID. PID records, at a minimum, describe the data in terms of its location (such as a URL) and fingerprint information, to facilitate integrity checks and access. These PIDs can then be associated for example to metadata.

The current data registry landscape is fragmented with many different offerings. The DONA-guided Handle System and its instantiations, such as DOI and EPIC, might be suitable for some disciplines to uniquely identify data (and other digital objects such as scientific workflows) as the basic infrastructure and tooling are already in place. In other cases, specialised data registries serve their domain well. Steps now need to be taken towards a truly scalable registry system that is open to all

[4] https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_pdf

interested disciplines worldwide and that enables data repositories to contribute to an open ecosystem for the long-term identification of scientific data. Ideally, such a worldwide system might federate the curation of datasets and enable ad-hoc sharing and collation of metadata for practical purposes.

	CRISP	ENVRI	DASISH	BioMedBridges
Data identity	■	■	■	■
Software identity	■	■	■	■
Data continuum	■	■	■	■
Concept identity		■	■	■
User identity management	■	■	■	■
Common Attribute Scheme	■	■	■	■
Common data standards and	■	■	■	■
Service discovery	■	■	■	■
Service market places	■	■	■	■
Integrated data access and	■	■	■	■
Data storage facilities	■	■	■	■
Data curation	■	■	■	■
Privacy and security	■	■	■	■
Dynamic data management		■	■	
Semantic annotation and bridging		■	■	■
User Community Body	■	■	■	■
Reference models	■	■	■	■
Education & training	■	■	■	■

Table 1 Data-related topics of common interest between the four cluster initiatives (shaded fields indicate interest)

Software identity

Just as with scientists and data, there is also a need for developers to publish their software and for other developers and researchers to access and cite software in a standard way. Traditional literature is not the ideal medium for this, especially for researchers who prefer developing rather then writing papers, and for software that is generally volatile. In addition, there is an increase in the amount of scientific data objects that are created automatically, by software as part of scientific workflows, rather than the traditional and manual human-driven process. For example, complex sensors may invoke a software process that performs many transformational operations, while complex scientific workflows may employ Web service components, which tend to be volatile, subject to updates and other changes.

Changes in complex computational systems make it difficult to understand exactly what software components and sensors were used to generate data objects. As a result, the science that relies on these computations suffers from a lack of repeatability and reproducibility. This is especially challenging where the software or an online service is developed using a continuous integration process without versions or releases. Similarly, ordered (orchestrated) software services are bundled, for example within workflow software, to define scientific analysis pipelines. Workflows themselves can also be seen as aggregations (here: sequentially executed software components) that need to be identifiable and citable.

Specific software components, workflow descriptions and other digital research objects[5] need to be identified by means of PIDs that resolve to metadata providing the attributes needed for practical applications[6], including for example, version information, contact details and documentation. Software citations should enable the attribution of credit to developers, help establish software provenance and promote sharing and reuse of software.

Data continuum

Scientific data is increasingly produced automatically and typically follows a data lifecycle. In this continuous process, new versions and objects are created, stored, registered and hence made referable (registered) and in some cases even citable. The latter includes a quality statement as the object (data) has been published by a researcher, thus endorsing its quality. This leads to a continuum (the data continuum) of objects from raw data to publications, all of which are referable or citable data objects.

An implication of the data lifecycle is that there is a need for different levels of visibility and persistency for the objects, including software and workflows, which may need to be referenced. Data must be linked in a way that ensures the continuum can be traversed. Working from the assumption that data is referenced using a PID, high-quality metadata to enable smart algorithms to ascertain the data flows, source-sink relationships, versioning etc., are required. Increasingly,

[5] http://eprints.ecs.soton.ac.uk/18555

[6] The working group “Associated Types” in the RDA initiative is investigating this problem

Virtual Collections (VC), which are identifiable and hence referable or citable objects, are being built to aggregate data objects for various purposes.

Concept identity

In many cases, scientific data (including metadata) is not only numeric but includes terms that convey semantics. To understand such data, the terms must be interpreted. Missing terms and inaccurate or ambiguous definitions of the concepts present a barrier to interpretation and hence re-use of the data in question.

In addition, semantics may shift over time. Semantic interoperability, already a difficult task, becomes impossible when the use of terms and semantics are not controlled. There is an urgent need for sustainable open concept registries (structured concept lists, thesauri, ontologies, controlled vocabularies etc.) that can be used by scientific communities to define, register and share their concepts.

Overall, semantic interoperability faces general technological, sociological and economic challenges. Technological challenges lie in the process of applying ontologies in data curation. The process is expensive, due in part to the complex nature of the work—dealing with variability, inconsistency and ambiguities in the original human descriptions and in the labour-intensive nature of performing curation. Even when curators are using ontologies, this does not guarantee inter-annotator agreement, as there is often an interpretation involved. Solutions to this, require appropriate tooling that can perform consistently across data, but are informed by human knowledge. The Zooma[7] tool, for example, aims to underpin the matching of sections of text to ontologies by repeating previous human-based assertions. It is a large knowledge base of 'curation rules' that allows past curation to be repeated consistently. Such approaches also offer the benefit of providing information on provenance, as to when an assertion was made and entered into the tool and by whom, offering an audit trail from original data to ontology-annotated data. This also enables updates of inconsistencies en masse in a consistent manner.

From a sociological stance, although ontologies and semantic descriptions have proven essential, for example in the life sciences, bioinformatics and humanities, an agreement on ontologies as 'standards' for a given domain can be divisive, sometimes causing branching and duplication of efforts. Accepting disagreements and capturing areas where people do not agree (semantically and explicitly) are currently avoided, as they are seen as representing a failure to reach consensus. However, such disagreements are often reflective of areas of science where consensus can also be hard to reach, and it may be worth capturing these areas in a more formal way. Finally, contributing to ontologies can be time consuming, and credit is often diluted such that the necessary effort is difficult to justify and ultimately avoided.

There are economic consequences from the general misconception, especially by funding bodies, that most of the ontologies that are required for semantic descriptions 'are built'. This is not the case: science continuously evolves and an ontology that becomes moribund after a three year

[7] http://www.ebi.ac.uk/fgpt/zooma/

project, will naturally become out of date, limiting its usefulness. The benefits of long-term sustainability are clear with projects such as the Gene Ontology[8].

The ISOcat data category registry[9] is one example from the humanities where the community started to register and explicitly define relevant concepts used to characterize phenomena of the different languages of the world. Since the languages spoken are so different, the semantics of the concepts describing them must also be different. However, being able to uniquely refer to a specific concept in an open registry is necessary to allow everyone to make use of them in their data. The availability of an exhaustive registry that people will use in their day-to-day work and which is supported by various tools is still far way. Although many other examples indicate that concept identity will be essential also for the humanities, progress will be slow.

User identity management

The amount of research data that is available and needs to be accessed by users is increasing rapidly. Open data that is accessible via the Internet, fosters re-usage, re-purposing, enrichment and creation of new data, not only by scientific researchers, but increasingly, by anyone who would like to contribute to the scientific effort, such as the citizen scientist. To ensure trust in the scientific output produced, it is not only necessary to identify the data, software and processes used, but also who created and used those objects in the data continuum. As data access crosses national and organisational borders, there is a need for interoperable systems for registration of the identities of the actors involved (user identity management).

In order to coordinate user identities internationally, a number of known obstacles must be removed: an agreed list of minimum user-related attributes (e.g. email address and home institute) that are needed by service providers must be agreed by identity federations at national level (e.g., the UK federation (UK), InCommon (US), SWITCHaai (Switzerland), SURFfederatie (Netherlands), etc. and internationally (e.g. eduGAIN). National federations must then encourage the release of this minimum set of attributes by Identity Providers at institutional level in each country. The user authentication and attributes must be reliable and up to date to ensure that the community data service providers can trust them, up to a certain level of assurance. In addition, there must be trusted attribute providers (such as REMS[10]) that the data service providers can build on, adding information on fine-grained access rights of the users, to the user-related attributes that are supplied by the identity federations. This aggregated information is passed to Service Providers. Note however, that there are still obstacles arising from requirements of national and EU data protection law on release and transmission of personally identifiable information that need to be overcome.

Common attribute scheme

A key challenge in research data management is the access management of increasingly large and valuable datasets. In some cases, access to data needs to be controlled e.g. in the case of sensitive or

[8] http://www.geneontology.org/
[9] http://www.isocat.org
[10] https://tnc2013.terena.org/core/presentation/18

personally identifiable data or data underlying certain restrictions, such as embargoes or copyrights, to ensure that the data in question is used only for the intended purpose and, again in the case of e.g. personally identifiable data, appropriate consent is available.

It is becoming apparent that there are huge advantages in terms of efficiency, security and trust in using federated identity management (FIM) vs. local log-on portals that need to be maintained for each dataset separately. With FIM and a common User Identity, service providers can refrain from using local log-on portals and instead delegate authentication to so-called identity providers or their agents. However, for more fine-grained access control, proper authorisation is still required on the service (or data) provider side. In this case, authorization is based on attributes attached to the digital (user) identity. Currently, attribute values are neither always consistently populated, or homogeneous. Hence, a commonly agreed scheme of attributes with widely and consistently defined values is needed.

Common data standards and formats

Common data standards and formats are required to allow data to be shared widely and to enable reuse or repurposing of existing software tools, without costly modifications. In order to share data (and software), either for validation or repurposing, that data needs to be understood in terms of its syntax (format/structure), meaning (concepts/semantics) and, ideally, the context in which it was generated (provenance).

Use of different standards and formats is a major barrier to data sharing. In addition, demand for tooling to support the standards and formats increases, creating a vicious cycle and resulting in a big loss of efficiency. The promotion of common standards and formats is therefore required for data and metadata descriptions, including provenance.

Both "top-down" and "bottom-up" (grass-roots) approaches must be used to encourage diverse scientific communities to use common data and metadata standards. There are several successful examples: in genomics, widely accepted standards have emerged organically, while in biology, the Proteomics Standards Initiative[11] has managed to rationalise data sharing and access. Top-down and bottom-up approaches are complementary: bottom-up input must be secured for discipline-specific items, and support and encouragement from the top down must be given to start making data interoperable over a wider range of scientific disciplines. Ideally—as the application of certain standards is spread over increasingly wide areas - there will eventually be fewer and fewer interfaces in the system.

Several disciplines have had successes in the establishment of widely accepted syntactic standards for specific data types. Often this is achieved through the activities of influential organisations, but in areas where no such de facto candidates exist, a specific collaborative standardisation initiative is required.

[11] http://www.psidev.info/

The fact that a worldwide, consistent and comprehensive solution to register models, schemas and (complex) scientific data types and formats is missing, makes crosswalks and re-use enormously difficult. For example, it is almost impossible for an occasional user who found a useful document of an unknown type, to quickly visualize the content.

Service discovery

Access to e-infrastructure has become indispensable for scientific research. Since e-infrastructure is composed of many independent services that may be managed by autonomous service providers, the ability of the user to discover suitable services - services that will enable them to conduct their research - can be difficult. Discoverability of services within and across different e-infrastructures is imperative and a precondition for their utilisation.

A coherent and comprehensive approach is required to achieve visibility of tools and services within the specialised domains and eventually, across disciplines and countries. In addition, a feasibility study to assess the potential of re-usable services would be highly useful. In any case, there is great potential for knowledge exchange across the disciplines and the possibility for technical exchange, e.g. sharing of winning strategies for building registries, sharing code and sharing metadata, if this is shown to be useful. Finally, registries may also be the starting point to support workflow orchestration, including across discipline boundaries. All of this is absolutely predicated upon registries that are sustainable, which can best be achieved by federating the curation burden amongst the community of service providers.

BioMedBridges/ELIXIR-DK are building a comprehensive tools and data service registry for the life sciences[12], including controlled vocabularies in support of consistent resource discovery[13]. DASISH has taken up the metadata component schema that was extended to services within CLARIN, where metadata is being harvested, made available via the Virtual Language Observatory and used for workflow orchestration, and is now working on a joint registry and portal for tools and services with enhanced functions.

Service marketplaces

Given the diverse landscape of e-infrastructure services and providers, virtual "marketplaces" are needed where users can compare offerings and select the best service provider, also in relation to any costs when imposed. While service registries are the foundation of such a marketplace, user commenting, experience documentation etc. are also required. Such transparent marketplaces will not only enable consumers to select the service that is optimal for their needs, but will also demonstrate the potential demand to the service providers, which will ultimately aid in setting development and support priorities.

Integrated data access and discovery

In order to share data across scientific disciplines, either for validation or repurposing, it is necessary initially, to be able to discover that data and gain access to it. This need brings together

[12] BioMedBridges tools and data service registry: http://wwwdev.ebi.ac.uk/fgpt/toolsui/

[13] http://bioinformatics.oxfordjournals.org/content/29/10/1325

many other requirements, such as metadata catalogues for data discovery, PIDs for unique identification of data, the data continuum so the original data that produces a certain result can be located, and a common user identity mechanism that is linked to transport protocols providing access to the data. The internal use of standards that provide the points of integration between different data sources, is a key aspect to providing layered, distributed integration of data. Communities interested in integrated data access and discovery must agree on such standards.

To achieve semantic interoperability and expose a given data landscape for discovery and easy use, a variety of different distributed integration technologies can be used: (1) REST-based "vignette" integration, which allows presentation of information from specific databases in a human readable form (these resources allow other websites to "embed" live data links with key information into other websites); (2) Web service based "query" integration, where simple object queries across distributed information resources can be used to explore a set of linked objects using dictionaries and ontologies; (3) scalable semantic Web-based technology, with data being exposed using e.g. RDF and SPARQL. The listing reflects a hierarchy where the lowest levels are the semantically poorest, but easiest to implement, whereas the highest levels potentially expose all information in databases that is both, permitted for integration and can be described using common standards.

Data storage facilities

The need to store data is a fundamental scientific requirement: data must be stored (even temporarily) so it may be processed and knowledge extracted. Data storage is needed for large data factories, such as high-energy physics and photon sciences or molecular biology on one end, as well as the small data producer such as individual scientists—the *long tail* of science—on the other. While there are commonalities in the requirements of actors at both ends of this spectrum, the types of storage facilities needed to meet data storage requirements differ. The driver of these differences is not the scientific domain alone, but the individual scientific processes involved.

Challenges with data storage services for small data producers, which are provided by various projects and commercial providers, include a certain lack of trust, presence, or absence of guarantees, for persistence and preservation of the uploaded data, and final costs.

Although both the physics and bioinformatics communities have many years of experience in handling very large datasets, both are facing new challenges. One of the biggest recent breakthroughs in the life sciences has been the development of high-throughput DNA sequencing technologies. The massive amount of molecular data now being produced is an emerging challenge for the storage infrastructures and it is clear that not one infrastructure alone can solve the problem. Similarly, the Physics and Astronomy community are experiencing rapid developments of instruments and detectors that generate extremely high data rates. The challenge of high-speed data recording is under investigation within CRISP alongside the broader challenge of data acquisition. The cost-effective storage and archiving of the resulting data volumes becomes an increasingly complex and challenging task, especially in situations where real-time data reduction is not an option.

In the social sciences and humanities, new challenges are emerging due to new crowd sourcing experiments that will engage thousands of participants, resulting in data streams of up to 1 TB/day/experiment. New ways of storing and pre-processing this data need to be put in place.

Data curation

All scientific data has intrinsic value, not only for the advancement of knowledge but also economically, based on the financial investments in the science that generated it. To gain maximum return on this investment, data must be made accessible to the wider research community through effective curation, with necessary metadata allowing the data to be understood and used. Without data curation, a huge amount of data may simply be archived and never exploited. This problem grows, as the amount of data generated explodes and data use intensifies. Open Access to scientific (and other) data is becoming increasingly topical and important. It is a golden opportunity for inter-disciplinary collaboration and for multi-disciplinary infrastructure(s) and services.

Data curation requires highly skilled specialists qualified in their scientific fields, who understand the data and can use supporting software. This essential role is currently significantly undervalued, with career progression heavily weighted towards traditional research activities.

There must also be policies to encourage researchers to deposit pre-curated data. In the many years of experience the biomedical sciences (bioinformatics) have in providing public archives for scientific data, it has proved to be extremely hard to persuade data submitters to provide even the most basic information on their data. Similar experiences have been gained by institutions in the social sciences and humanities domain in the case of large archives or the aggregation of information about objects. To ensure that submitters do provide sufficient information on their data, policies are required that encourage researchers and projects to deposit them and associated "knowledge" (software, metadata, documentation etc.). Deposition of research data in a suitable format and including all necessary metadata and provenance information must be a key part of the research process.

Privacy and security

The handling of data with ethical, legal and societal implications (ELSI) has become a challenge in the biomedical sciences as well as the social sciences and humanities. Legal and societal aspects are increasingly being faced within domains such as e.g. genetically modified organisms, measuring the energy consumption of a private household, recording personal behavior and parameters, or earthquake predictions. It is therefore critical that appropriate e-secure systems to store and provide ELSI data are developed.

As an example, in the field of genomics research, the current practice is that data access committees within the respective database/institute handle requests to use sensitive data, and the access applications are submitted by each individual researcher or research group for each individual dataset.

This process of individually assessing each request will not scale for the era of genomics. Infrastructure to manage this in an automatic yet secure way is urgently required. This could be

implemented, for example, via a model where ELSI data would be stored and remain in a trusted repository using a fully-controlled cloud storage, with access to the data being restricted to certified researchers that use certified software.

It is crucial that ELSI data remains available to be used in research, and it must be a priority to ensure that there are means to securely provide access to sensitive data.

Dynamic data

Dynamic Data, such as environmental real-time measurements, can be characterised by a continuously changing content of the Digital Objects (DO) they consist of, which is caused by asynchronous processes. Dynamic Data sets are mutable and dependent on asynchronous processes during data acquisition. Dynamic Data needs to be part of the registered domain of data, i.e. it must be referable and citable, replicated to guarantee persistence etc. This is also important in the context of quality assessment and control of data streams.

When planning data management processes, the special characteristics of datasets need to be considered. Among these are the PIDs that should be associated with data streams; another issue is dealing with streaming analytics of parallel data streams. At a point in time during measurements and streaming, depending on the performance of the data-producing equipment (such as sensors in remote areas or mobile devices used for large-scale crowd sourcing), datasets may be incomplete. They may also include systematic errors due to the production context. In contrast to immutable data, where version registration is controlled by explicit steps from humans or within workflow chains, mechanisms for identification, replication etc. have not been established in the case of Dynamic Data.

There is an urgent need for a Data Fabric solution for Dynamic Data, which is an environment where automatic workflows (based on widely agreed policies and practices) curate generated data in real-time so that its Digital Objects can be managed, cited, accessed and used in all phases.

Semantic annotation and bridging

Data can originate from many independent sources, each of which may have its own semantics. It is important for interdisciplinary researchers to have a methodology by which information from a large number of sources can be associated, organised, and merged. Semantic annotation can be seen as a common service that can be applied to processes of data enrichment and quality assurance in many scientific disciplines. Semantic annotation and anchoring of data to ontologies increases the feasibility of semantic bridging, which is a paradigm that is becoming increasingly important.

Disparate sources of data can automatically be related via overarching ontologies. It is possible to interrelate any pair of ontologies indirectly through semantic bridges consisting of many other previously unrelated ontologies, even when there is no way to determine a direct relationship between them. The relationships among the ontology fragments indicate the relationships among the sources, enabling the source information to be categorised and organised.

Tools exist that allow researchers to annotate data against ontologies and to map free-text annotations to trusted ontologies. This is especially important for (semi-) automatic services that curate data in order to manage systematic errors, incomplete metadata, or adaptation to changing metadata concepts.

Supporting tasks

User community body

The relevance of an e-infrastructure is defined by its users and the scientific research that it supports. The requirements of the researcher must therefore drive the continuous development of any e-infrastructure for it to remain relevant. As the importance of European e-infrastructures grows and matures, it becomes increasingly important that user communities are able to voice requirements and help drive the direction of their evolution. Similarly, the European e-infrastructure providers themselves need to understand the requirements of a wide variety of user communities, which in general, are not necessarily the end users themselves, but the institute or project that supports them. For individual research infrastructures, as well as cooperating infrastructures in a scientific domain, it would be beneficial to exchange experiences in consultation processes with the user communities and vice versa, where mechanisms allow users to raise their ideas and suggestions in an organised way, or even influence the direction of infrastructure development.

EIROforum has proposed[14] a pan-European user forum for organisations and projects that operate at an international level, in order to present to the policy makers and the infrastructure providers where there are common needs and opinions, and where there is divergence. This will provide both policy makers and e-infrastructure providers with a view across many research domains, enabling them to take strategic decisions that will reflect the commonalities, and differences, that exist.

It should be noted that many initiatives have applied many different strategies to optimise user engagement. The proposed pan-European forum should be seen as a complementary and necessary layer to optimise engagement, as existing user interaction programs of the e-infrastructures remain highly fragmented.

Reference models

A Reference Model defines a uniform framework, against which an infrastructure's components can be classified and compared, providing a common language for communication. It can help to identify common solutions to similar problems, enabling the reuse of resources and the sharing of experiences and thus avoiding duplication of efforts. Reference Models are based on standard descriptions of data, computation and research infrastructure services and consequently, provide authority and stability. The cooperating research infrastructures in ENVRI are developing a common Reference Model which proved to be beneficial to enhance their services and to promote their infrastructure interoperability, both of which are crucial for the growing number of multi-disciplinary projects addressing the grand challenges in environmental research.

[14] http://cds.cern.ch/record/1545615/

The adoption of a common Open Distributed Processing (ODP) framework for communication, not only provides a unified view across different scientific domains, but also raises awareness for areas that require further attention.

Education and training

Education and training are key to the uptake of e-infrastructure. Operators (technicians, service providers, finance, and management) need to understand how best to deliver the advantages of e-infrastructure and communicate to the end user how their work will benefit.

Even concerning the topics covered in this paper there is huge variability of knowledge and expertise within and between the cluster projects. Such variability can hamper collaboration and make communication on certain topics—or even the identification of such topics—extremely difficult, jeopardising opportunities for knowledge exchange and the identification of possible synergies.

Overall, the cluster projects should share experiences and identify best practice in the education and training of end-users in the use of Research Infrastructures. This knowledge should then be widely disseminated across the communities covered by the cluster projects, to inform the wider community about the different (and possibly new) approaches for consuming and providing services. This could be achieved by joint knowledge exchange workshops, organised by the cluster with the most expertise in the area in question, and possibly through activities where collaborating Research Infrastructures exchange staff on a time-limited basis.