European Strategy Forum
on Research Infrastructures

e-IRG Report on Data Management

Data Management Task Force

November 2009

1 EXECUTIVE SUMMARY

A fundamental paradigm shift known as Data Intensive Science is quietly changing the way science and research in most disciplines is being conducted. While the unprecedented capacities of new research instruments and the massive computing capacities needed to handle their outputs occupy the headlines, the growing importance and changing role of data is rarely noticed. Indeed it seems the only hints to this ever burgeoning issue are mentions of heights of hypothetical stacks of DVDs when illustrating massive amounts of "raw, passive fuel" for science. However, a shift from a more traditional methodology to Data Intensive Science – also sometimes recognized as the 4th Research Paradigm – is happening in most scientific areas and making data an active component in the process. This shift is also subtly changing how most research is planned, conducted, communicated and evaluated. This new paradigm is based on access and analysis of large amounts of new and existing data. This data can be the result of work of multiple groups of researchers, working concurrently or independently without any partnership to the researchers that originally gathered the information.

Use of data by unknown parties for purposes that were not initially anticipated creates a number of new challenges related to overall data management. Long-term storage, curation and certification of the data are just the tip of the iceberg. So called **Digital Data Deluge**, for example, caused by the ease with which large quantities of new data can be created, becomes much more difficult to deal with in this new environment. Large amounts of data are created not only by state-of-the-art scientific instruments and computers, but also by processing and collating existing archived data.

These challenges have been discussed in length during the 2007-2008 e-IRG workshops, and particular focus has been put on data initiatives linked with the new research infrastructures identified by ESFRI. The e-IRG delegates recognised the importance of data management for the future of research infrastructures and, as a result, established the e-IRG Data Management Task Force (DMTF) that received recognition and support also from ESFRI. The main objectives of DMTF were defined as producing an analysis of issues regarding data management in a coherent and flexible way, and a set of recommendations for the present and future research infrastructures. DMTF, following the e-IRG instructions, put together a large group of experts in data management who have prepared the present report.

The following report is divided into three parts: a survey of existing data management initiatives, metadata and quality, and interoperability issues in data management. *The report ends with conclusions of the study and a set of proposed recommendations for further analysis and discussion by the e-IRG*. The findings will also be presented to the e-IRG and to ESFRI to create a final set of recommendations endorsed by the two bodies.

This work is part of a continuous analysis of the scientific data situation, started initially with various contributions from the European Commission (communication published in February 2007, OECD position about scientific data status, published in xx 2007 and ESFRI position paper about Digital Repositories in September 2007). At the same time the DMTF was elaborating the present contribution, the European Commission issued also a quite ambitious strategic paper about the increasing role of e-Science. All together, these various contributions will help defining the needs and the terms of a Global Data Infrastructure, relevant to participate to the Lisbon strategy in building the European Research Area.

The present report addresses the following points in further detail:

1) Existing data management initiatives includes an initial inventory of known data management initiatives from operational initiatives to future projects. This survey analyses the opportunities, synergies and gaps presented by these initiatives or clusters of similar initiatives and their potential impact. The survey is divided into three main fields of science: arts and humanities – social sciences; health sciences; and natural sciences and engineering. For each scientific field a large number of initiatives is analysed regarding e.g. data type, standards used or proposed, data curation, quality control, and best practices for preservation. Challenges and needs relevant to various user communities are also identified. The analysis of 18 social science, 12 health sciences and 33 natural sciences and engineering initiatives gives a global view of most of the data initiatives in Europe. Achieving this comprehensive coverage necessitated a trade-off in the form of leaving a detailed study of collaboration opportunities and synergies (both between e-IRG and the initiatives and between initiatives) to be accomplished by future initiatives. Furthermore, due to the very diverse and rapidly evolving nature of data initiatives and the links between them, the study indicates that e-IRG should carefully focus its efforts to the most promising cases when establish contacts or actively facilitating collaboration between the initiatives and communities. This is necessary in order to maximise

the impact that can be achieved with reasonable resource investments.

2) Metadata and quality covers the basic principles and requirements for metadata descriptions and the quality of the resources to be stored in accessible repositories. The principles and requirements specified are considered to be baselines for all research infrastructures, and as such are independent of the scientific research field. Key findings of this part of the report focus on metadata flexibility to allow for addition of new elements, for using different types of selections and for the possibility of using elements of different sets and re-using existing elements/sets. Metadata topics such as usage, scope, provenance, persistence, aggregation, standardisation, interoperability, quality, earliness and availability are discussed in detail. Quality of data resources is also covered in the context of sharing data and quality assurance, assessing the quality of research data, and data consumers.

3) Interoperability issues in data management are very important to ensure that scientific data is reachable and useful to other scientific fields, i.e. to enable cross-disciplinary Data Intensive Science. The same data is often relevant to researchers in different scientific fields, but it is not always obvious that researchers can access and use data outside their own domain with the same ease and efficiency as discipline-based solutions. Thus interoperability is fundamental for allowing flexible and coherent cross-disciplinary data access. At the moment interoperability-related activities are mainly contained within individual communities, but, with the advent of e-Science, data interoperability needs to be extended to groups of different communities. In addition to providing details of these opportunities and challenges, this part of the document presents several levels and types of interoperability: resource-level operability, general semantic operability and syntactic versus semantic interoperability. The different layers, ranging from device level to communications, middleware and deployment of resource interoperability are analysed in detail. Semantic interoperability is also discussed in terms of data integration, ontology support, simplicity, transcoding and metadata, representation information, conceptual modelling, and distributed systems. Some use cases in the medical field, linguistics, e-humanities ecosystem, earth sciences, astronomy and space science, and particle physics are presented and, in each instance, use cases solutions, tendencies and needs are identified and put forward.

This report is obviously not intended as the final word in the area of data management, rather it aims to put together several starting points to encourage future efforts in this domain. The participating authors sincerely hope that interested parties will take on board the findings of this document and craft them into a group of concrete, well-aligned initiatives, which fulfil the promises of Data Intensive Research. The authors also wish that the e-IRG and ESFRI, when drafting recommendations based on these findings will, for their part, ensure that these initiatives have clear and efficient contact points with e-Infrastructure policy makers in the future!

2 Summary of Recommendations

All recommendations and proposed guidelines are embedded in the full texts of the various annexes. However a short synthesis of them is summarized as follows:

2.1 Metadata

R1: Usage Providing metadata describing any kind of research resources and services is an urgent requirement for service providers and resource repositories.

R2: Scope There is an increasing pressure for disciplines to agree on a set of semantically specific enough elements that allows researchers to describe their services and resources.

R3: Provenance Descriptive metadata should include or refer to provenance information to support long-term preservation and further processing.

R4: Persistence Metadata descriptions need to be persistent, to be identified by persistent identifiers and also to refer to the resources and services they represent by using persistent identifiers.

R5: Aggregations Descriptive metadata have an enormous potential to describe various forms of groupings and can give them an identity, i.e. making them citeable.

R6: Standardization Descriptive metadata needs to be based on well-defined element semantics and a schema-based format to cater for presentations for humans and machine operations. Where fixed schema solutions are given up, elements need to be re-used which are registered in open registries.

R7: Interoperability Descriptive metadata needs to be open and offered for harvesting via widely accepted mechanisms to cater for interdisciplinary usage.

R8: Quality Researchers need to be urged to produce high quality metadata descriptions.

R9: Earliness Researchers should be motivated to create metadata immediately and tool developers should add those descriptors that can be created automatically.

R10: Availability It is a MUST for all resource and service providers to create and provide quality metadata descriptions.

2.2 QUALITY

Recommendations for quality of data are the synthesis of various contributions such as OECD, ICSU and RIN and various other sources. The reader will find all details and references in Chapter 2.

2.3 INTEROPERABILITY

R11 Actively encourage programmes that support cross-disciplinary access to digital objects and related services.

R12 Encourage the development of non-discipline-specific frameworks and information architectures for interoperable exchange of data;

R13 Support communities for the definition of their requirements and their activities in the domain of semantic interoperability;

R14 Support interoperation activities within multinational *and* multi-disciplinary / community grids, e.g. OGF activities, or within EGI; the activity itself, however, is likely to be focused on a part of the infrastructure, e.g. authentication, job submission, or storage;

R15 Prioritise those interoperation activities aiming at standardising interfaces and/or protocols, or documenting current usage and limitations of existing standards for interfaces and protocols;

R16 Ensure that work is practical and realistic instead of theoretical "paperwork";

R17 Ensure that besides hardware and services, digital objects deserve infrastructure components in their own right:

- mediation services for metadata / semantic annotations of data;
- persistent linkage of research data with publications and other resources;
- policies for long-term preservation of data, maybe focused into dedicated centers (preservation activities plus consultation);

R18 Define proper governance structures and guidelines for (inter)national agreements for distributed heterogeneous data facilities;

R19 Highlight that the basis of proper data management is a proper repository setup with strict organizational guidelines that are supported as widely as possible by a proper repository system;

R20 Highlight that for achieving semantic interoperability in open scenarios the project oriented approach of formal ontologies seems to be problematic, suggesting that a separation between concept definitions and their relations is desirable (as is suggested by ISO 11179 and ISO 12620 for example).

R21 Highlight that state-of-the-art network infrastructures are needed that are capable of adapting flexibly to the needs of the applications and researchers relying on them.

3 Task Force Membership

3.1 Survey

Patrick Aerts, Andreas Aschenbrenner, Lalos Balint, Hilary Beedham, Victor Castelo, Brian Coghlan, Ana Bela Sa Dias, Rudolf Dimper, Luigi Fusco, Francoise Genova, David Giaretta, Jonathan Giddy, Matti Heikkurinen, Maria Koutrokoi, Michèle Landes, Carlos Morais-Pires, Christian Ohmann, Pasquale Pagano, Leonard Rivier, Lorenza Saracco, Dany Vandromme, Peter Wittenburg and Hans Zandbelt.

3.2 Metadata and Quality

Patrick Aerts, Hilary Beedham, Tobias Blanke, Victor Castelo, Peter Doorn, Luigi Fusco, David Giaretta, Matti Heikkurinen, Maria Koutrokoi, Michèle Landes, Diego Lopez, Pasquale Pagano, Dany Vandromme, Peter Wittenburg and Andrew Woolf.

3.3 Interoperability

Patrick Aerts, Andreas Aschenbrenner, Hilary Beedham, Brian Coghlan, David Corney, Rudolf Dimper, Luigi Fusco, Françoise Genova, David Giaretta, Matti Heikkurinen, Jens Jensen, Maria Koutrokoi, Wolfgang Kuchinke, Michèle Landes, Diego Lopez, David O'Callaghan, Christian Ohmann, Pasquale Pagano, Jonathan Tedds, Miroslav Tuma, Dany Vandromme, Peter Wittenburg, Andrew Woolf and Hans Zandbelt.

Contents

Chapter 1

Survey of projects and initiatives in the field

1 EXECUTIVE SUMMARY

This chapter presents an inventory of existing European data management initiatives. It covers operational initiatives as well as future projects. The initial objectives of analysing the opportunities, synergies and gaps between these initiatives, as well as an analysis of the potential impact on the e-IRG, have not been reached. It has also not been possible within the available time frame to contact the data initiatives in order to verify the information contained in this report and to start informal discussions in view of initiating collaboration opportunities. These ambitious aims are left to a future initiative, which would however have to target such actions to a small number of specific areas to be sure to obtain a tangible result within a reasonable amount of time.

The following general statements can be concluded from this survey:

- A very large number of individual and domain specific data initiatives and scientific databases exist in Europe,
- Only a small number of the data initiatives are federated,
- Interoperability is de facto limited to a scientific domain which has applied standard formats in an early stage of the data initiative,
- Long-term sustainability is a major issue for all data initiatives, and not only a problem limited to the underlying hardware infrastructure but also for the software accessing and exploiting the data,
- Open access to data has not yet become a reality in all scientific domains. It is often technically hampered by the absence of search engines, and institutionally by the absence of clear policy guidelines,
- Organizations are moving away from a centralised stand-alone model towards distributed networks of federated data repositories,
- New requirements for cross disciplinary research will require interoperability between different disciplines and different types of data,
- Metadata is recognised as being paramount for long-term data access and usability (including documenting the research process, not just the data itself),
- The suite of data analysis tools are growing and becoming more complex (i.e. the use of GIS in many fields),
- The focus is on curating data for reuse, not necessarily for long-term preservation,
- Open access to data is becoming more common,

- New projects, many of them financed by the EU, will profoundly change the current data landscape in Europe and set standards for the rest of the world,
- Additional efforts should be put in gathering more detailed information about the project specific needs on data management and what kind of activities are required,
- Communication and cooperation between the data initiatives/projects should be stimulated to achieve better interoperability and reuse of solutions and infrastructures.

2 Introduction

Today, research is increasingly data intensive and relies on access to large and complex data sets. Unprecedented access to data and digital tools is increasing the efficiency of research and enabling new discoveries. Indeed, the effectiveness of research depends on how data is managed. In the past data archiving has been managed at the level of the discipline, community, or individual researcher. However, given the substantial cost of creating data collections and the complexity of managing and preserving them, this approach is often no longer considered adequate.

2.1 Purpose

This document aims at a survey of existing data management initiatives for consideration by the e-IRG. The intended audience for this document include the e-IRG delegates and the e-IRG Data Management Task Force (DMTF).

2.2 Scope

There are thousands of data collections around the world. The intent of this survey is by no way to provide an exhaustive list of data initiatives, but instead to identify a set of data initiatives in broad scientific disciplines. Because this is a rapidly evolving area, a list of "demonstrator" and "in development" projects that were encountered during the study are also included in the report.

Within the time-frame and the resources available for this study it has not been possible to contact the various data initiatives nor to provide an estimate of the potential impact on the e-IRG.

2.3 Organization, Methodology

This review provides a scan of European data initiatives, identified through the existing literature and Internet searches.

Data archiving initiatives are sorted in three broad disciplinary categories: (1) Arts and Humanities, Social Sciences, (2) Health Sciences, and (3) Natural Sciences and Engineering.

2.4 Overview

Digital data comes in many forms and ranges from raw data to highly processed data and results finally in publications. Data management activities differ significantly according to data types and the mandate of the organisations involved. The types of data collected by an individual data archive are very discipline specific. While there are a few archives that collect a large variety of data types within a single archive (such as DANS in the Netherlands), most others have been specifically designed to collect and manage a particular data types in a given discipline.

The scope of data archives differ widely. National archives in the social sciences tend to collect data of that individual nation, while scientific archives are more likely to collect data beyond national borders for their community or discipline. Some archives collect raw data, others focus on post-analysis data only, and some collect both.

Most repositories adhere to discipline specific metadata standards. In the social sciences, data centres regularly employ the Data Documentation Initiative (DDI) standard. No general model for the representation of scientific metadata exists. However, there is a nascent movement to develop a common set of metadata so that datasets from different scientific disciplines are interoperable, and frequently the Dublin core metadata initiative (DCMI)[64] can be found as the basis for such metadata definitions.

The aim of many data archives is to curate data in order to allow further analysis beyond the original experiment or measurement. For many of the data initiatives it is not clear what specific preservation measures are applied by the organisation. Many archives refer to best practices, such as the Open Archival Information System reference model (OAIS). Costs, integrity and reproducibility of data are considerations for preservation.

Data is acquired in a number of ways:

- In the Natural Sciences, data is generally acquired directly from scientific instruments such as telescopes, satellites, or synchrotrons
- In the UK, the government-run data centres do require researcher deposit
- In the Arts and Social Sciences, it is usually a combination of researcher deposit and acquisition (sometimes by payment) of external data sets from other organizations
- In the Health Sciences, where there is a tradition of journals requiring data deposit before publishing related articles, the responsibility often lies with the researcher to deposit the data

3 Survey of Data Initiatives

3.1 Arts and Humanities, Social Sciences

In social sciences, important datasets are often collected not by research teams but by government departments to inform policy-makers. These data are usually of a high quality, are national samples or census data, and are often under-utilised with high secondary value as a source of information for the social science research. The social science archives recognised this potential early and some have been able to negotiate access to these data for the wider research community. This is not always the case and the reformed CESSDA will result in an extension of access to such data. SSH (social science and humanities) research data collected for research funded by research councils and other funders such as the Commission are normally only shared beyond the award-holding team at the time or very soon after, they have published their results. In some countries plans for sharing are required as part of the original award (in UK for example, ESRC funded researchers must contact UKDA at the beginning of their research and have to offer their data for archiving on completion of the award. Sanctions are applied if this is not done.) This practice is expected to become more widespread across Europe and is under discussion, with CESSDA contributing, at DG research (for the social sciences).

CASPAR - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval

CASPAR [19] is an integrated project co-financed by the EU within the Sixth Framework Programme and intends to provide tools and techniques for secure, reliable and cost-effective preservation of digitally encoded information for the indefinite future. CASPAR claims compliance to the OAIS reference model and intends to enhance techniques for capturing representation information and other preservation related information for content objects, technology independence. It integrates digital rights management, authentication, and accreditation. Design virtualisation services are supposed to allow for long term digital resource preservation despite changes in the underlying computing and storage systems.

CLARIN - Common Language Resources and Technology Infrastructure

The CLARIN [20] project is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily usable. CLARIN offers scholars the tools to allow computer-aided language processing, addressing one or more of the multiple roles language plays (i.e. carrier of cultural content and knowledge, instrument of communication, component of identity and object of study) in the Humanities and Social Sciences. The CLARIN initiative offers a comprehensive service to the humanities disciplines with respect to language resources and technology, tools and resources that will be interoperable across languages and domains, thus addressing the issue of preserving and supporting the multilingual and multicultural European heritage, and a persistent and stable infrastructure that researchers can rely on for the next decades.

CLARIN will be built on and contribute to a number of key technologies coming from the major initiatives advancing the eScience paradigm, like the data Grid technology to connect the repositories, semantic Web technology to overcome the structural and semantic encoding problems, and advanced multi-lingual language processing technology that supports cultural and linguistic integration.

TextGrid

TextGrid [22] aims to create a community grid for the collaborative editing, annotation, analysis and publication of specialist texts. It thus forms a cornerstone in the emerging e-Humanities. Building on existing expertise in the field of e-Science and advancing towards the Semantic Grid, TextGrid partners are developing a comprehensive toolset for researchers in philology, linguistics, and related fields. Reaching out to the academic community, the project establishes an interdisciplinary platform for research. Open interfaces open the door for other projects to plug into the TextGrid. TextGrid is part of the D-Grid initiative, and is funded by the German Federal Ministry of Education and Research (BMBF). With a project start in 2005, TextGrid still has funding through 2011 and is looking for ways to ensure sustainability with the growing TextGrid community and inter/national partners.

IASSIST - International Association for Social Science Information Service and Technology

IASSIST [81] is an international organization of professionals working with information technology and data services to support research and teaching in the social sciences. Its 300+ members work in a variety of settings, including data archives, statistical agencies, research centres, libraries, academic departments, government departments, and non-profit organizations. IASSIST has the following objectives:

- To encourage and support local and national information centres for social science data.
- To foster international exchange and dissemination of information regarding substantive and technical developments related to social science data.
- To coordinate international programs, projects, and general efforts that provide a forum for discussion of issues relating to social science data.
- To promote the development of standards for social science data.
- To encourage educational experiences for personnel engaged in work related to these objectives.

DOBES - Dokumentation Bedrohter Sprachen

The DOBES [23] programme aims at documenting languages that are potentially in danger of becoming extinct within a few years time. Since 2000, 50 documentation projects have been funded and there will be calls for concrete documentation projects until 2011. From the beginning the DOBES programme wanted to take advantage of modern state-of-the-art technology, and where necessary drive technology to suit the needs of the documentation work. Therefore, the following topics were discussed and widely agreed upon, in particular in the pilot phase:

- specifications for archival document formats to promote long-term accessibility

- recommendations for recording and analysis formats, and tools to ensure quality and reduce the conversion effort
- the creation of new tools that support the audio/video annotation work, the metadata creation and the navigation in metadata domains, advanced web-based frameworks to access and enrich archived resources.

HRELP Hans Raising Endangered Languages Project

HRELP [24] aims to document endangered languages, train language documenters, preserve and disseminate documentation materials, and support endangered languages. It operates from a donation of £20 million from Arcadia. HRELP is based at SOAS, University of London, and consists of three programmes:

- The Documentation Programme (ELDP) is providing £15 million in research grants to document the world's most endangered languages
- The Academic Programme (ELAP) Teaches postgraduate courses in language documentation and description, and field linguistics. It also hosts post-doctoral fellows, researchers, visitors, and conducts seminars and training
- The Archiving Programme (ELAR) is preserving and disseminating endangered language documentation, developing resources, and conducting training in documentation and archiving

DARIAH - Digital Research Infrastructure for the Arts and Humanities

DARIAH's [25] mission is to support digitally enabled research in the arts and humanities. DARIAH aims to develop and maintain an infrastructure in support of ICT-based research practices across the arts and humanities, acting as a trusted intermediary between disciplines and domains. DARIAH is working with communities of practice to:

- Develop and apply ICT-based methods and tools to enable new research questions to be asked and old questions to be posed in new ways
- Link distributed digital source materials of many kinds
- Provide access to digital research collections
- Exchange knowledge, expertise, methodologies and practices across domains and disciplines

The research that DARIAH supports will expand the knowledge and understanding of our heritage, histories, languages and cultures. DARIAH is part of the ESFRI Roadmap and thus equipped to establish a sustainable infrastructure. While its preparatory phase has only begun in late 2008, it is working with numerous international partners (e.g. Bamboo http://projectbamboo.org/, Interedition http://interedition.huygensinstituut.nl/) to achieve these goals.

CESSDA - Council of European Social Science Data Archives

Background

The social Sciences and humanities have a long history of data sharing. CESSDA [26] has existed as an informal organisation, promoting data sharing, for over 30 years. The organisation currently comprises 20 data archives in as many countries and collectively serve some 30,000+ social science and humanities researchers and students within the European Research Area each year, providing access to 25,000 data collections, delivering over 70,000 data collections per annum and acquiring a further 1,000 data collections each year. CESSDA has long supported data-sharing in the social sciences and has operated a trans-border data exchange agreement for many years and has a history of encouraging the deposit of data for secondary analysis from both academic researchers and official producers of data such as government departments and ministries. With support from successive EU contracts, CESSDA has also worked on technical aspects of data management, by developing tools and web-based

technologies to facilitate data discovery, access and dissemination and by active participation in the development and application of appropriate standards for interoperability and data harmonisation. Developments include the CESSDA data portal, the ELSST multilingual thesaurus, used for resource discovery and a significant contribution from a number of the member organisations, to the development of the DDI standard and, more recently, the Data Seal of Approval. National CESSDA members are currently funded differentially with the result that activity and commitment is greater in the better funded countries. This is now being addressed, since CESSDA was recognised a European Research Infrastructure, following receipt of PPP (Preparatory Phase Project) funding from the EU.

CESSDA-ERIC

From January 2008, CESSDA has received EC funding to reform the organisation from an informal grouping to an ERIC (Education Resources Information Centre) which is expected to be inaugurated in April 2010. The new organisation will place CESSDA on a secure financial footing and enable it to extend the benefits of membership to more member states and to non-European organisations with similar goals for data sharing in the social sciences and humanities. CESSDA-ERIC will co-ordinate European data archiving and sharing for the social sciences and humanities. It will set standards for membership and promote data sharing by:

- Setting standards for data-management activities (e.g. metadata standards for data, archival standards)
- Knowledge transfer for potential new members of CESSDA, with the goal of widening access to data
- Working with international data producers to negotiate access to data for research purposes
- Working with organisations that fund research (nationally and pan-European) to promote and facilitate the sharing of publicly funded data
- Co-ordinating training in data management (for both data producers wishing to share data and archive staff)
- Providing data support services for data archiving and sharing
- Providing a Shibboleth-based single sign on/authentication service, enabling researchers in any member country to access data in any other country without the requirement for multiple access requests
- Developing tools to facilitate the inclusion of catalogue records and data into the CESSDA data portal
- Extending the ELSST thesaurus to include new languages (currently it includes 9)
- Monitoring new technologies to ensure the best possible services for users

Social science data

CESSDA recognised that the diversity of data sources presented a problem of resource discovery for researchers, particularly those wishing to undertake cross-national research. In response, it developed the CESSDA data catalogue, with its multilingual thesaurus, as an aid to resource discovery. The CESSDA Catalogue enables users to locate datasets, as well as questions or variables within datasets, stored at CESSDA archives throughout Europe. Data collections include sociological surveys, election studies, longitudinal studies, opinion polls, and census data. Among the materials are international and European data such as the European Social Survey, the Eurobarometers, and the International Social Survey Programme. A number of CESSDA members have worked to break down the barriers to data sharing within the research community: data confidentiality and protection; concerns that the complexity of data would make it difficult for them to be understood outside the original research team; concerns that other teams may publish ahead of the original research team and; financial costs and additional effort of preparing data for sharing. Most of these issues are not unique to the SSH community and some members, for example, the UKDA, collaborate with other disciplines, e.g. in the fields of medicine, environment and biological and biomedical sciences, to promote data sharing in other communities and also cross-disciplinary access to data.

Structure of services

CESSDA's governance structure has been designed to maximise flexibility of service provision to researchers. It is fundamentally a distributed organisation but permits the provision of both centralised or distributed services at national level so that Country A may wish to provide ingest, preservation and dissemination services from

a single organisation whilst Country B may wish to offer each of these functions from different organisations. Similarly, the new structure will include a membership category for data producing organisations or projects which have established their own data management and distribution systems but have shared interests with CESSDA. Collaboration CESSDA-ERIC is collaborating closely with other ESFRI SSH projects. Representatives of the five projects meet to discuss common problems and representatives of CLARIN and DARIAH are invited to CESSDA meetings with common interests (and vice versa).

Lifewatch is also considering problems associated with metadata that reside separately from the data. The general tradition in social science has been to store the data and metadata together, with the archive working with the depositor (or data collector) to collect or generate (using tools), the metadata needed to make a dataset reusable. However, there are some datasets for which there are country specific versions and each archive will have metadata for their national data plus, sometimes, metadata linking to the data in other countries. For this and other reasons, CESSDA has a group looking at persistent identifiers, edition and version control. CESSDA is prototyping a shibboleth based system for authentication and registration across several organisations in different countries.

DANS - Data Archiving and Networked Services

DANS [21] is an institute under the auspices of the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NWO) and is a member of several ESFRI Roadmap initiatives (among which CESSDA, DARIAH and CLARIN). Since its establishment in 2005, DANS has been storing and making research data in the arts and humanities and social sciences permanently accessible. It develops permanent archiving services, stimulates others to follow suit, and works closely with data managers to ensure as much data as possible is made freely available for use in scientific research. To permanently store and access data, the location where the data is stored is much less important than the way it is stored. To ensure that archived data can still be accessed, recognised and used in the future, DANS developed a DATA hallmark, which now internationally is recognized as the Data Seal of Approval (http://www.datasealofapproval.org, see also AppendixA). This hallmark can be requested and granted to research data repositories that meets a number of clear criteria in the field of quality, permanence and accessibility of the data and provides the research financiers with the guarantee that research results remain accessible for reuse.

UKDA - UK Data Archive, University of Essex

The UK Data Archive (UKDA) is a centre of expertise in data acquisition, preservation, dissemination and promotion and is curator of the largest collection of digital data in the social sciences and humanities in the UK. It is funded by the Economic and Social Research Council (ESRC), the Joint Information Systems Committee (JISC) of the Higher Education Funding Councils and the University of Essex. Founded in 1967, it now houses several thousand datasets of interest to researchers in all sectors and from many different disciplines. The UKDA currently holds the Presidency of CESSDA and is co-ordinator of the CESSDA PPP. UKDA is also a member of the International Association of Social Science Information Service and Technology (IASSIST) through which it plays a lead role in international collaborative projects on issues such as data sharing, metadata and social science thesauri and is a member institution of the US national social science and historical data archive, Inter-university Consortium for Political and Social Research (ICPSR) in Michigan and the International Federation of Data Organizations (IFDO). The UKDA provides resource discovery and support for secondary use of quantitative and qualitative data in research, teaching and learning. As a lead partner of the Economic and Social Data Service (ESDS), the UKDA is responsible for:

- overall integration and management of the ESDS
- access and preservation, focusing on the central activities of data acquisition, processing, preservation and dissemination
- ESDS Qualidata, a specialist service for a range of qualitative datasets
- ESDS Longitudinal, undertaken jointly with the Institute for Social and Economic Research (ISER)

and supports:

- ESDS International, working with Manchester Information and Associated Services (MIMAS), providing access to international micro data
- ESDS Government, working with the Cathie Marsh Centre for Census and Survey Research (CCSR), facilitating access to large-scale government datasets

The UKDA also provides preservation services for other data organisations, supports the National Centre for e-Social Science (NCeSS) and facilitates international data exchange through agreements with other national archives. The UKDA hosts the History Data Service and Census.ac.uk, facilitating access to the census data resources for UK higher and further education. The UKDA's involvement in Research and Development projects has made a significant contribution to new developments in data preservation and dissemination, metadata standards, software for web browsing, data discovery and data delivery. The UKDA also hosts the Rural Economy and Land Use Programme (RELU) Data Support Service (DSS) and the Secure Data Service which will provide secure access to sensitive data. This service will become operational in Autumn 2009.

ESDS - Economic and Social Data Service

The Economic and Social Data Service [29] is a national data archiving and dissemination service which came into operation in January 2003. The service is a jointly-funded initiative sponsored by the Economic and Social Research Council (ESRC) and the Joint Information Systems Committee (JISC).

The ESDS is a distributed service, based on a collaboration between four key centres of expertise:

- UK Data Archive (UKDA), University of Essex
- Institute for Social and Economic Research (ISER) , University of Essex
- Manchester Information and Associated Services (MIMAS), University of Manchester
- Cathie Marsh Centre for Census and Survey Research (CCSR), University of Manchester

These centres work collaboratively to provide preservation, dissemination, user support and training for an extensive range of key economic and social data, both quantitative and qualitative, spanning many disciplines and themes. The ESDS provides an integrated service offering enhanced support for the secondary use of data across the research, learning and teaching communities.

The overall direction and management for ESDS is the responsibility of the UK Data Archive (UKDA), providing consistency and standards across the service. It performs a broad strategic role, relating to a variety of stakeholders concerned with the supply, funding and use of social science data, and creates a coherent publicity, promotion and outreach strategy for the whole service.

ESDS Management supports high quality research, teaching, and learning in the social sciences by acquiring, developing and managing social and economic data and related digital resources; and by promoting, disseminating and supporting the use of these resources as effectively as possible.

The service provides procedures for data preparation, preservation, processing and documentation, cataloguing and conditions of use, and has developed and maintained common standards across ESDS, ensuring interoperability between ESDS and other services providers is achieved and maintained.

Under the UK Data Protection Act the ESDS has a legal duty to protect any information it collects from users. The ESDS Privacy policy relates to personal data collected by the ESDS in the course of registration for access to services and during user consultations.

RELU-DSS - Rural and Economy and Land Use Data Support Service

RELU-DSS is hosted at the UKDA to oversee and implement the RELU data management policy and to support RELU award holders in developing and achieving their data management plan. RELU projects investigate the social, economic, environmental and technological challenges faced by rural areas in the UK in an interdisciplinary manner. Teams of social and environmental researchers study topics such as restoring the public's trust in food chains, tackling animal and plant disease in a socially acceptable manner, enabling sustainable farming

in a globalised market, developing land management techniques to deal with climate change and managing land and water use for sustainable water catchments. Central to this research is the integration of social, economical, biological, agricultural and environmental science data. RELU-DSS offers an interdisciplinary data support service, co-ordinated between the UK Data Archive (UKDA), a service provider for the Economic and Social Data Service (ESDS), based at the University of Essex, and the Centre for Ecology and Hydrology (CEH) at Lancaster, through its Environmental Informatics Programme. RELU-DSS provides information and guidance to researchers and project managers on data management, data sharing and preservation. Research data resulting from RELU projects will be archived and preserved at the UK Data Archive, the data centre of the Economic and Social Research Council (ESRC), and at CEH Data Centres for the Natural Environment Research Council (NERC). They will be preserved long-term as digital data and made available to the research community for wider use. Advice and guidance for RELU award holders is provided through a telephone and email service, web-based guidance and workshops for researchers on data management, legal and ethical issues and the practicalities of depositing data for preservation and archiving. RELU-DSS also informs and makes recommendations to the RELU management bodies on key data management issues. RELU is funded by ESRC, NERC and Biotechnology and Biological Sciences Research Council (BBSRC).

SHARE - Survey of Health, Ageing and Retirement in Europe

SHARE [27] is the upgrade to a lasting infrastructure of a multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social and family networks of more than 40 000 individuals aged 50 or over. SHARE is coordinated centrally at the Mannheim Research Institute for the Economics of Aging (MEA). It is harmonized with the U.S. Health and Retirement Study (HRS) and the English Longitudinal Study of Ageing (ELSA). SHARE's scientific power is based on its panel design that grasps the dynamic character of the ageing process. SHARE's multi-disciplinary approach delivers the full picture of the ageing process. Rigorous procedural guidelines and programs ensure an ex-ante harmonized cross-national design.

Data collected include health variables (e.g. self-reported health, health conditions, physical and cognitive functioning, health behaviour, use of health care facilities), bio-markers (e.g. grip strength, body-mass index, peak flow), psychological variables (e.g. psychological health, well-being, life satisfaction), economic variables (current work activity, job characteristics, opportunities to work past retirement age, sources and composition of current income, wealth and consumption, housing, education), and social support variables (e.g. assistance within families, transfers of income and assets, social networks, volunteer activities). In addition, the SHARE data base features anchoring vignettes from the COMPARE project and variables and indicators created by the AMANDA RTD-Project. The data are available to the entire research community at no cost.

COMPARE

The COMPARE [28] project collects survey data aimed at creating internationally comparable measures of several dimensions of the quality of life - health, economic position, work disability, contacts with family and friends, health care quality, political efficacy, and satisfaction with life as a whole. Its full name is "Toolbox for Improving the Comparability of Cross-National Survey Data with Applications to SHARE".

Quality of life is typically measured with subjective questions, such as "how good is your health?" with possible answers "very good", "good", "moderate", "fair" or "poor." International comparisons of such answers may be biased if people in different countries use these response scales differently. For example, someone in Denmark may answer his health is "very good," but someone with exactly the same health but living in Portugal may answer "good." COMPARE aims at making answers comparable by correcting for such response scale differences. The method that is used for this is the technique of Anchoring Vignettes.

Anchoring vignettes are short descriptions of, e.g., the health or job characteristics of hypothetical persons. Respondents are asked to evaluate the hypothetical persons on the same scale on which they assess their own health or job. Respondents are thus providing an anchor, which fixes their own health assessment to a predetermined health status or job characteristic. These anchors can then be used to make subjective assessments comparable across countries and socio-economic groups.

COMPARE is part of the family of research projects linked to SHARE, the Survey of Health, Ageing and Retirement in Europe. Data collection is in parallel to the SHARE data collection in waves 2004 and 2006-2007

and follows the same procedures. The sample covers respondents of age 50 and older and their spouses in 10 EU countries: Sweden, Germany, Poland, Netherlands, Belgium, France, Czech Republic, Spain, Italy and Greece. Similar data will also be collected in Denmark (funded by NIA). In all other countries data collection is funded by the European Commission through the STREP project COMPARE # 028857 in the Citizens and Governance in a Knowledge-Based Society Programme.

DCC - Digital Curation Centre

The Vision of the DCC [30] is to be the

- Centre of excellence in digital curation and preservation in the UK
- Authoritative source of advocacy and expert advice and guidance to the community
- Key facilitator of an informed research community with established collaborative networks of digital curators
- Service provider of a wide range of resources, software, tools and support services

This is acknowledged as an ambitious vision but one which is built on the strong foundations achieved during the first phase, and which will enable the DCC to act as an agent of transformational change to facilitate digital curation best practice within the rapidly changing environment of e-Research.

The DCC has the following objectives:

1. Provide strategic leadership in digital curation and preservation for the UK research community, with particular emphasis on science data
2. Influence and inform national and international policy
3. Provide advocacy and expert advice and guidance to practitioners and funding bodies
4. Create, manage and develop an outstanding suite of resources and tools
5. Raise the level of awareness and expertise amongst data creators and curators, and other individuals with a curation role
6. Strengthen community curation networks and collaborative partnerships
7. Continue our strong association with our research programme

The DCC is funded by JISC and the e-Science Core Programme.

DPC - Digital Preservation Coalition

The DPC [31] is a not-for profit membership organisation whose primary objective is to raise awareness of the importance of the preservation of digital material and the attendant strategic, cultural and technological issues. It acts as an enabling and agenda-setting body within the digital preservation world and works to meet this objective through a number of high level goals. Its vision is to make our digital memory accessible tomorrow.

In order to achieve this aim, the Coalition has the following long-term goals:

- producing, providing, and disseminating information on current research and practice and building expertise amongst its members to accelerate their learning and generally widen the pool of professionals skilled in digital preservation.
- Instituting a concerted and co-ordinated effort to get digital preservation on the agenda of key stakeholders in terms that they will understand and find persuasive.
- Acting in concert to make arguments for appropriate and adequate funding to secure the nation's investment in digital resources and ensure an enduring global digital memory.

- Providing a common forum for the development and co-ordination of digital preservation strategies in the UK and placing them within an international context.
- Promoting and developing services, technology, and standards for digital preservation.
- Forging strategic alliances with relevant agencies nationally and internationally, and working collaboratively together and with industry and research organisations, to address shared challenges in digital preservation.
- Attracting funding to the Coalition to support achievement of its goals and programmes.

ICPSR - Inter-University Consortium for Political and Social Research

Established in 1962, ICPSR [32] is the world's largest archive of digital social science data. ICPSR acquires, preserves, and distributes original research data and provides training in its analysis. ICPSR also offers access to publications based on their data holdings.

NCeSS - National Centre for e-Social Science

The NCeSS [33] is funded by the Economic and Social Research Council (ESRC) to investigate how innovative and powerful computer-based infrastructure and tools developed over the past five years under the UK e-Science programme can benefit the social science research community. This infrastructure is commonly known as the 'Grid'.

e-Social Science refers to the use of Grid infrastructure and tools within the social sciences. The role of NCeSS is to investigate specific applications of e-Social Science, develop tools to support them and to advise on the future strategic direction of e-Social Science and e-Science. NCeSS also provides information, training, advice, support and online resources to help the social science research community adopt e-Social Science.

The centre consists of a coordinating Hub at the University of Manchester, seven Research Nodes, and 12 Small Grant projects. However, this structure will change from October 2009 when the hub will cease to continue. The research programme includes applications of e-Science in both quantitative and qualitative social sciences, and studies of issues relevant to promoting the wider adoption of e-Science. A series of smaller e-social science projects have been commissioned under the ESRC Small Grant project scheme.

3.2 Health Sciences

The European Union's Member States are committed to sharing their best practices and experiences to create a European eHealth Area, thereby improving access to and quality health care at the same time as stimulating growth in this industrial sector. The European eHealth Action Plan [85] plays a fundamental role in the European Union's strategy. Work on this initiative involves a collaborative approach among several parts of the Commission services. The European Institute for Health Records [86] is involved in the promotion of high quality electronic health record systems in the European Union.

An important consideration in the process of developing electronic health records is to plan for the long-term preservation and storage of these records. The field will need to come to consensus on the length of time to store EHRs, methods to ensure the future accessibility and compatibility of archived data with yet-to-be developed retrieval systems, and how to ensure the physical and virtual security of the archives.

Additionally, considerations about long-term storage of electronic health records are complicated by the possibility that the records might one day be used longitudinally and integrated across sites of care. Records have the potential to be created, used, edited, and viewed by multiple independent entities. These entities include, but are not limited to, primary care physicians, hospitals, insurance companies, and patients. Mandl et al. have noted that "choices about the structure and ownership of these records will have profound impact on the accessibility and privacy of patient information."[87]

The required length of storage of an individual electronic health record will depend on national and state regulations, which are subject to change over time. Ruotsalainen and Manning have found that the typical preservation time of patient data varies between 20 and 100 years. In one example of how an EHR archive might function, their

research "describes a co-operative trusted notary archive (TNA) which receives health data from different EHR-systems, stores data together with associated meta-information for long periods and distributes EHR-data objects. TNA can store objects in XML-format and prove the integrity of stored data with the help of event records, timestamps and archive e-signatures."[88]

While it is currently unknown precisely how long EHRs will be preserved, it is certain that length of time will exceed the average shelf-life of paper records. The evolution of technology is such that the programs and systems used to input information will likely not be available to a user who desires to examine archived data. One proposed solution to the challenge of long-term accessibility and usability of data by future systems is to standardize information fields in a time-invariant way, such as with the XML language.

ELIXIR - European Life-Science Infrastructure for biological Information

The mission of ELIXIR [34] is to construct and operate a sustainable infrastructure for biological information in Europe to support life science research and its translation to medicine and the environment, the bio-industries and society.

The proposed infrastructure will ensure free provision of essential biological data to the entire scientific community. It will encompass an interlined collection of robust and well-structure and evaluated core databases, capable of accommodating the ongoing massive accumulation and diversification of data. It will permit the integration and interoperability of diverse, heterogeneous, potentially redundant information that is essential to generate and utilise biomedical knowledge. It will encompass the necessary major computer infrastructure to store and organise this data in a way suitable for rapid search and access, and will provide a sophisticated but user-friendly portal for users. It will be embedded in a database-related research programme that supports the development of critically important standards, ontologies and novel information resources.

The ELIXIR home page provides access to a compilation of more than 500 databases from ELIXIR affiliated countries as well as a geographical distribution of these databases.

EATRIS - European advanced translational research infrastructure in medicine

EATRIS [35] is a strategic EU project that aims to offer a research infrastructure to help overcome bottlenecks currently hampering the transfer both of basic research findings into clinical application and of clinical observations to basic research. In a unique partnership, governmental and scientific organisations form the EATRIS consortium to develop a master plan for setting up the provision of an infrastructure on a European level. The EATRIS idea is to organize under one roof multidisciplinary, creative work atmosphere, open labs, comprehensive modern equipment, scientific and legal expertise with central facilities and services and a translational research curriculum.

ECRIN - European Clinical Research Infrastructures Network

ECRIN [36] is the pan-European Infrastructure for clinical trials providing, high-quality services to multinational clinical research.

As a distributed infrastructure linking national networks of clinical research centres and clinical trials units, ECRIN provides integrated 'one-stop shop' services to investigators and sponsors in multinational studies, with the local contribution of staff embedded in each national coordination. Such support is particularly relevant for academic clinical research, research on rare diseases, neglected diseases, and for clinical trials sponsored by biotechnology, drug, and device enterprises that may face difficulties in conducting multinational studies in the EU. ECRIN will be a major tool for FP7- or Innovative Medicines Initiative (IMI)-funded projects. Hereby ECRIN will stimulate EU research on prevention, diagnosis and treatment, hence improving healthcare delivery to patients and citizens.

ECRIN is designed to bridge the fragmentation of clinical research in Europe through the interconnection of national networks of clinical research centres and clinical trial units. ECRIN plans extension to national infrastructure networks in other member states, and stimulates the set-up of new national networks for further connection to ECRIN. This integrated clinical research infrastructure, unique in the EU, provides support to any type of clinical research, and in any medical field.

A first step (ECRIN - Reciprocal Knowledge Programme, 2004-2005, FP6 Health priority) helped identify bottlenecks to multinational cooperation and define the strategy for future development. In the second step (ECRIN-TWG, 2006-2008, FP6 Health priority), procedures and guidelines for multinational studies in the EU were prepared by transnational working groups. In the third step (ECRIN-PPI, 2008-2011, FP7 Infrastructure Unit) the European infrastructure for clinical trials further develops and provides services to pilot multinational clinical research.

Finally ECRIN establishes contacts in other world regions to promote connection with regional clinical research infrastructures worldwide, with the objective of sharing best practices and developing interoperability.

In the "comparative analyses on clinical research infrastructures, networks, and their environment in Europe" [16] report one can read:

There is a major diversity in national rules and regulations regarding data management (e.g. archiving, data protection). Procedures and tools used for data management differ widely between ECRIN members (e.g. SOPs, software). In the majority of centres no professional commercial software is routinely used except for data-analysis. Only exceptionally data management audits have been performed. Experiences are available in individual centres/networks of ECRIN for systematic software-evaluation, use of professional commercial software and data management audits.

A major problem in the ECRIN network is the missing harmonization of procedures and quality management related to data management. The situation is complicated by limited financial resources, high prices for commercial software and uncertainty in the software marketplace. So far validated software is not used on a regular basis. Standards such as CDISC, MedDRA are rarely applied. Integration between study software tools and clinical information systems has not been performed.

A primary task is to improve quality management in data management (e.g. harmonized SOPs). The suitability of software products for academic research should be evaluated (e.g. Open Source software). The processes and quality management of data management should be harmonized. Professional study software should be implemented and validated. Interfacing and integration should be supported by standards. Data management audits should be supported. A further task for the future would be the implementation of centralized data bases for research.

EU-OPENSCREEN - European Infrastructure of Open Screening Platforms for Chemical Biology

The EU-OPENSCREEN [37] facility will allow researchers in academia and SMEs to access resources for the development of bioactive small molecules. It will be an association of high throughput screening (HTS) centres. These offer chemical resources for hit discovery and optimisation, bio- and cheminformatics support, and a publicly accessible database. This database combines screening results, assay protocols, and chemical information. A central facility will make available a large collection of diverse compounds representing the chemical knowledge of Europe.

Background:

Chemical Biology, the use of small organic molecules to explore biology, provides unique means for unravelling complex biological processes. As a major goal, Chemical Biology aims to identify small-molecule modulators for individual functions of proteins. As tools for academic research, these will enable a deeper exploitation of the wealth of genomic information. The efficacy and impact of the approach depends largely on the availability of a diverse and well-designed compound collection, the most advanced screening technologies, chemistry resources, special cell line collections, bio- and chem-informatics capacities, and a comprehensive database.

Impact foreseen

For the first time, European researchers from academia and SMEs will obtain access to the most advanced screening technologies. This will allow the researchers to identify compounds affecting new targets. The interdisciplinary approach of EU-OPENSCREEN will bring together chemists, engineers, informaticians and biologists, overcoming the fragmentation of European research in the field of Chemical Biology. Through EU-OPENSCREEN´s coordinated and transnational activities, a substantially accelerated generation of knowledge will be achieved. In particular as regards the responses of biological systems challenged by small molecules. EU-OPENSCREEN aims to satisfy the needs for new bioactive compounds in many fields of the Life Sciences (e.g. human and veterinary medicine, systems biology, biotechnology, agriculture and nutrition).

EU-OPENSCREEN will primarily support projects on unconventional targets and that address fundamental biological questions. In this respect, the activities of EU-OPENSCREEN will precede commercial development. They will open new paths for research in the post-genomic era, and a more direct translation from basic science into an improved quality of life. EU-OPENSCREEN will be open to all European organisations involved in Chemical Biology and committed to open access. In order to facilitate both collaboration amongst members and their interaction with stakeholders and external users, it is envisaged to create a legal entity.

A flexible framework for Intellectual Property (IP) issues will be established to allow for an early protection of knowledge before publishing in the database. Thus, the necessary balance between rapid knowledge sharing and exploitation activities will be secured.

INFRAFRONTIER - European infrastructure for phenotyping and archiving of model mammalian genomes

INFRAFRONTIER [38] will organize two complementary and linked infrastructure networks for large-scale and comprehensive phenotyping and archiving of mouse models serving the European genetics and biomedical research community for the benefit of human health.

Central objective of the preparatory phase is to commit all relevant stakeholders to actively participate in the joint development of Infrafrontier. In addition to the participating scientific partners this involves in particular funding agencies and ministries of the European member states.

Infrafrontier integrates 15 European laboratories with exceptional track records to implement and run large-scale infrastructures. Infrafrontier builds on existing infrastructures under EMMA and EUMODIC and forms a coalition with a significant number of funding agencies to develop the prerequisites to a common European infrastructure.

The Infrafrontier preparatory phase aims to organize a stable and sustainable infrastructure by

- the identification of the most suitable legal form,
- developing a business plan based on a sustainable funding concept,
- reaching a legal agreement between all partners and
- providing a strategic plan for the construction phase.

Infrafrontier will therefore give Europe a leading position in the worldwide competition on resources and knowledge for medically relevant mouse models by providing a user-driven platform.

BBMRI Biobanking and Biomolecular Resources Research Infrastructure

BBMRI [39] is a pan-European and internationally broadly accessible research infrastructure and a network of existing and de novo biobanks and biomolecular resources. The infrastructure will include samples from patients and healthy persons, representing different European populations (with links to epidemiological and health care information), molecular genomic resources and biocomputational tools to optimally exploit this resource for global biomedical research.

BBMRI will be composed of a network of centres organized in a distributed hub structure comprising:

- biobanks of different formats (collections of blood, DNA, tissue, etc., together with medical, environmental, life-style and follow-up data),
- biomolecular resources (antibody and affinity binder collections, ORF clone collections, siRNA libraries, proteins, cellular resources etc.),
- enabling technologies and high-throughput analysis platforms and molecular tools to decipher gene, protein and metabolite functions and their interactions,
- harmonized standards for sample collection, storage, pre-analytics and analysis,
- harmonized databases and biocomputing infrastructure and

- ethical, legal and societal guidance platform.

The Structure of BBMRI

Key components of BBMRI are comprehensive collections of biological samples from different (sub-) populations of Europe, which should be linked with continuously updated data on the health status, lifestyle and environmental exposure of the sample donors. This can only be achieved in a federated network of centres established in most, if not all, European Member States. Therefore, the format of BBMRI should be a distributed hub structure in which the hubs coordinate activities, including collection, management, distribution and analysis of samples and data for the major domains. The biobanks, biomolecular resources and technology centres, which are members of BBMRI, are associated with their specific domain hub. Furthermore, a variety of public or private partners (e.g., universities, hospitals, companies), which provide biological samples, data, technologies or services, may be associated with certain BBMRI members.

- BBMRI members represent the key providers of resources and technologies. Members are leaders in the field and drivers of innovation and scientific excellence. Membership is non-exclusive so that members link BBMRI to other national, European (e.g., other FP7 programs) and global initiatives (e.g., the emerging OECD global network of Biological Resource Centres or WHO programmes).
- Associated partners and subcontractors provide certain resources (services, data, samples, materials) to BBMRI. An associated partner, for instance, a hospital or research institute which provides biological samples and data, may be either reimbursed or compensated for its contribution by being granted free access to resources and technologies of the BBMRI. Associated partners may also be ministries, governments, research councils, and funding agencies from interested countries whether or not they currently support biobank or biomolecular resource infrastructure projects
- Users may come from different fields of academia and industry. Access will be provided in the context of specific research projects and on the basis of secured funding. Incentives may be provided for EU Member States and for industry to enter into general user agreements.

This structure provides great flexibility so that new members and partners can be connected at any time and so that it can be adapted to emerging needs in biomedical research. The IT infrastructure which employs federated database architecture and grid computing technology will integrate the complex network of hubs, members and partners. Hubs will be coordinated and directed by an executive management, which is supported by a governance council as well as by a high-calibre advisory board and receives input from the stakeholder forum to guarantee clear responsibilities as well as open and transparent decision-finding processes.

BBMRI will link to several ongoing international activities, such as those pursued by P3G, the Innovative Medicines Initiative, ISBER, the OECD, and the WHO, as well as research projects funded under FP5/FP6 and new projects under FP7. To avoid duplication of activities, BBMRI will exchange concepts and experience with these activities.

EBI - European Bioinformatics Institute

EBI [40] is a non-profit academic organisation that forms part of the European Molecular Biology Laboratory (EMBL).

As we move towards understanding biology at the systems level, access to large data sets of many different types has become crucial. Technologies such as genome-sequencing, microarrays, proteomics and structural genomics have provided 'parts lists' for many living organisms, and researchers are now focusing on how the individual components fit together to build systems. The hope is that scientists will be able to translate their new insights into improving the quality of life for everyone. However, the high-throughput revolution also threatens to drown us in data. There is an ongoing, and growing, need to collect, store and curate all this information in ways that allow its efficient retrieval and exploitation. The European Bioinformatics Institute (EMBL-EBI), which is part of the European Molecular Biology Laboratory (EMBL), is one of the few places in the world that has the resources and expertise to fulfil this important task. The Institute manages databases of biological data including nucleic acid, protein sequences and macromolecular structures.

The EBI Mission is:

- To provide freely available data and bioinformatics services to all facets of the scientific community in ways that promote scientific progress
- To contribute to the advancement of biology through basic investigator-driven research in bioinformatics
- To provide advanced bioinformatics training to scientists at all levels, from PhD students to independent investigators
- To help disseminate cutting-edge technologies to industry

BioSapiens

The objective of the BIOSAPIENS [41] Network of Excellence is to provide a large-scale, concerted effort to annotate genome data by laboratories distributed around Europe, using both informatics tools and input from experimentalists.

The Network will create a European Virtual Institute for Genome Annotation, bringing together many of the best laboratories in Europe. The institute will help to improve bioinformatics research in Europe, by providing a focus for annotation and by the organisation of European meetings and workshops to encourage cooperation, rather than duplication of effort.

An important aspect of the network activities is to try and achieve closer integration between experimentalists and bioinformaticians, through a directed programme of genome analysis, focused on specific biological problems. The annotations generated by the Institute will be available in the public domain and easily accessible on the web. This will be achieved initially through a distributed annotation system (DAS), which will evolve to take advantage of new developments in the GRID.

EMBRACE - European Model for Bioinformatics Research and Community Education

EMBRACE [42] will standardise access to bioinformatics resources, enabling data providers to provide well-defined interfaces to their databases that will conform to the same standards. This will allow users to make the most of dispersed data resources.

The objective of EMBRACE is to draw together a wide group of experts throughout Europe who are involved in the use of information technology in the biomolecular sciences. The EMBRACE Network of Excellence will optimise informatics and information exploitation by pure and applied biological scientists in both the academic and commercial sectors.

The network will work to integrate the major databases and software tools in bioinformatics, using existing methods and emerging Grid service technologies. The integration efforts will be driven by an expanding set of test problems representing key issues for bioinformatics service providers and end-user biologists. As a result, groups throughout Europe will be able to use the EMBRACE service interfaces for their own local or proprietary data and tools.

EMMA - European Mouse Mutant Archive

EMMA [43] is a non-profit repository for the collection, archiving (via cryopreservation) and distribution of relevant mutant strains essential for basic biomedical research.The laboratory mouse is the most important mammalian model for studying genetic and multi-factorial diseases in man. Thus the work of EMMA will play a crucial role in exploiting the tremendous potential benefits to human health presented by the current research in mammalian genetics.

The EMMA network is a partnership of several laboratories and other institutions throughout Europe. The current membership includes the CNR Istituto di Biologia Cellulare in Monterotondo, Italy (core structure), the CNRS Centre de Distribution, de Typage et d'Archivage animal in Orleans, France, the MRC Mammalian Genetics Unit in Harwell, UK, the KI Karolinska Institutet in Stockholm, Sweden, the FCG Instituto Gulbenkian de Ciência in Oeiras, Portugal, the HMGU Institute of Experimental Genetics in Munich, Germany, the EMBL European Bioinformatics Institute in Hinxton, UK, the GIE-CERBM Institut Clinique de la Souris, Illkirch, France, the Wellcome Trust Sanger Institute in Hinxton, UK and the CSIC Centro Nacional de Biotecnología in Madrid,

Spain. The EMMA network is directed by Professor Martin Hrabé de Angelis who also heads the HMGU/IEG in Munich. To ensure the operation of such a large and international enterprise an effective management structure consisting of several components was implemented. EMMA is open for the incorporation of new members into the current network to share the increasing workload and guidelines for this process were established.

EMMA is supported by the partner institutions, national research programmes and by the EC's FP7 Capacities Specific Programme.

EUMODIC - European Mouse Disease Clinic

EUMODIC [44] will undertake a primary phenotype assessment of up to 650 mouse mutant lines derived from ES cells developed in the EUCOMM project. Lines showing an interesting phenotype will be subject to a more in depth assessment.

EUMODIC will build upon the comprehensive database of standardised phenotyping protocols, called EMPReSS, developed by the EUMORPHIA project. EUMODIC has developed a selection of these screens, called EMPReSSslim, to enable comprehensive, high throughput, primary, phenotyping of large numbers of mice.

Health-e-Child

The Health-e-Child [80] project aims at developing an integrated healthcare platform for European paediatrics, providing seamless integration of traditional and emerging sources of biomedical information. The long-term goal of the project is to provide uninhibited access to universal biomedical knowledge repositories for personalised and preventive healthcare, large-scale information-based biomedical research and training, and informed policy making.

The Health-e-Child project focus will be on individualised disease prevention, screening, early diagnosis, therapy and follow-up of paediatric heart diseases, inflammatory diseases, and brain tumours. The project will build a Grid-enabled European network of leading clinical centres that will share and annotate biomedical data, validate systems clinically, and diffuse clinical excellence across Europe by setting up new technologies, clinical workflows, and standards.

Objectives

- To gain a comprehensive view of a child's health by vertically integrating biomedical data, information, and knowledge, that spans the entire spectrum from genetic to clinical to epidemiological;
- To develop a biomedical information platform, supported by sophisticated and robust search, optimisation, and matching techniques for heterogeneous information, empowered by the Grid;
- To build enabling tools and services on top of the Health-e-Child platform, that will lead to innovative and better healthcare solutions in Europe:
- Integrated disease models exploiting all available information levels;
- Database-guided biomedical decision support systems provisioning novel clinical practices and personalised healthcare for children;
- Large-scale, cross-modality, and longitudinal information fusion and data mining for biomedical knowledge discovery.

3.3 Natural Sciences and Engineering

The natural sciences and engineering domain experience an unprecedented data avalanche which is fuelled to a large extend by the fast evolution of sensor/detector technology leading to much more powerful instruments. The advances in IT also contribute to the data avalanche by enabling to capture, analyse, and store unbelievable quantities of data. Advances in technology make it more difficult to distinguish between raw data and processed data because preprocessing takes place in or very close to the detectors to correct and normalise data. Metadata

describing each of the steps leading to the final result is crucial to permit scientists others than those who have done the experiment to reproduce the results or re-use the data for further analysis.

The necessity to preserve data in institutional repositories is not yet well accepted in all scientific domains, if only because of the strong competition for the resources this requires or because of the inherent competition between scientists. Curating, preserving and making data accessible requires continues encouragement from the funding bodies. This encouragement should lead to clear policies within the scientific domains, ideally even cross discipline to prepare and foster cross disciplinary research. The e-IRG has an important role to play in this area. It must be kept in mind that a lot of research projects concern unique samples which cannot be studied easily and/or which require very expensive examination methods. Some of these data sets are unique and extremely valuable for future preservation, be it for further reference or re-newed analysis with new and better software. Data mining applications will emerge as a new tool for scientific discovery for certain areas of science.

It became clear during the survey that there are quite different domain specific practices in the natural sciences. Astronomy and earth and environmental sciences are well organised communities with well established best practices to preserve their data. This is not the case for the High Energy Physics community [84] despite on-going efforts around the upcoming LHC. A nascent initiative around neutron and photon sources in Europe has started discussions how to preserve the data originating from the increasing number of these large facilities for multidisciplinary research. This initiative will try to address the need to preserve and structure the access to data generated by a community of more than 25000 scientists in Europe.

BODC - British Oceanographic Data Centre

BODC [45] has earned an international reputation for its expertise in the management of marine data. In the 1980s they pioneered an 'end to end' approach, working alongside marine scientists during the lifetime of projects to ensure good data management practices. As well as ensuring data quality they facilitate data exchange between project participants before delivering the final data set on CD-ROM. Today, funding bodies insist that all marine projects contain adequate provision for professional data management. BODC holds publicly accessible marine data collected using a variety of instruments and samplers and collated from many sources. The data holdings can be used for science, education and industry, as well as the wider public. BODC makes data available under a license agreement.

OpenDOAR - The Directory of Open Access Repositories

OpenDOAR [46] is an authoritative directory of academic open access repositories. Each OpenDOAR repository has been visited by project staff to check the information that is recorded here. This in-depth approach does not rely on automated analysis and gives a quality-controlled list of repositories.

As well as providing a simple repository list, OpenDOAR lets you search for repositories or search repository contents. Additionally, OpenDOAR provides tools and support to both repository administrators and service providers in sharing best practice and improving the quality of the repository infrastructure. Further explanation of these features is given in a project document Beyond the list.

The current directory lists repositories and allows breakdown and selection by a variety of criteria - see the Find page - which can also be viewed as statistical charts. The underlying database has been designed from the ground up to include in-depth information on each repository that can be used for search, analysis, or underpinning services like text-mining. The OpenDOAR service is being developed incrementally, developing the current service as new features are introduced. A list of Upgrades and Additions is available.

Developments will be of use both to users wishing to find original research papers and for service providers like search engines or alert services which need easy-to-use tools for developing tailored search services to suit specific user communities.

The importance and widespread support for the project can be seen in its funders, led by the Open Society Institute (OSI), along with the Joint Information Systems Committee (JISC), the Consortium of Research Libraries (CURL) and SPARCEurope.

OpenDOAR has also been identified as a key resource for the Open Access community (K.B.Oliver & R.Swain, 2006 - PDF) and was one of the services which contributed to SHERPA being awarded the SPARC Europe Award for Outstanding Achievements in Scholarly Communications.

DRIVER-II - Digital Repository Infrastructure Vision for European Research

DRIVER-II [47] establishes a network of relevant experts and Open Access repositories. DRIVER II is a project funded by the 7th Framework Programme of the EC. The DRIVER web site holds details on the state of development of repositories in different European countries and how other repositories can join this growing European network. DRIVER uses the OAI-PMH and OAI-DC (syntactical layer) and vocabularies (semantic layer).

The DRIVER project responds to the vision that any form of scientific-content resource should be freely accessible through simple Internet-based infrastructures. The project is a joint collaboration between ten international partners with the intention to create a knowledge base of European research. DRIVER will put a test-bed in place across Europe to assist the development of a knowledge infrastructure for the European Research Area.

The DRIVER project consists of eight work packages. In work package 7, a number of studies were carried out under the label 'focused studies':

- An inventory study into the current types and level of Digital Repository (DR) activity - described in this report.
- An inventory study of important DR-related issues and good practices, such as business models, IPR, long-term preservation/archiving/access, data curation and stimuli for depositing materials into DRs.
- An investigative study of technical standards.

The results of these focused studies will be used in the other work packages within the DRIVER project, especially in the work package 8 'Awareness-raising and Advocacy programme'.

The study is aimed at making a complete inventory of the current state of digital repositories in 25 countries of the European Union. It is a follow up of an earlier SURF study carried out in 2005, which covered 10 European countries. The study was started in June 2006 and completed in February 2007. A Digital Repository is in this study defined as: 1 Containing research output 2 Institutional or thematic 3 OAI compliant.

Within a single country, there are benefits to be gained from providing a level of coordinated national support for repository development. This support is of two main types: support for the establishment and development of repositories; and support for the use of repositories. The first is developmental, providing support structures for a national network of repositories to be established. The second is service-based, providing support for the use of repositories by academics and the maintenance of repositories by their administrators. This support is likely to be required in the longer term.

There are four principal models for national developmental support structures.

Collaborative partnerships This model builds on a partnership of institutions working together in a mutually supportive initiative. Partners benefit from peer-level support and the identification and sharing of best practice. Economies can be made by establishing shared technical resources and shared management. In some cases it might be appropriate to share a repository, at least in the initial stages. An example of this approach is the UK-based SHERPA Project and partnership.

National support programmes Where there is a national coordinating body which can operate across institutions, one option is to fund a national development programme, where institutions can get support and information from a single centralised point, that is relevant to their needs. This has the advantage of allowing each institution to proceed at its own rate, which providing sufficient centralised support to help promote working standards and share best practice. This requires political will and commitment from a central body, together with significant levels of funding. The UK-based RSP - Repositories Support Project is an example of this approach.

Comprehensive national partnerships Creating a fully comprehensive support partnership for all institutions is more suitable for a smaller country or one with fewer institutions. This approach brings the benefit of a coordinated national response, with the ability to set joint working practices, standards. As the partnership is working within a single legislative and administrative environment, there are advantages in administration and policy development. There are also economies of scale in providing technical and administrative support. However, with larger numbers of institutions involved, the complexity of working across institutions can mean that the model is impractical.

An example is the Dutch repositories model, underlined by the National Academic Research and Collaborations Information System NARCIS which provides access to thousands of scientific publications and data sets, as well as information on researchers (expertise), research projects and research institutes in the Netherlands.

Centralised repository services Studies have found that institutional repositories offer advantages of longevity and stability of access for materials, as well as giving various benefits to the institutions themselves. However, there are circumstances where centralised repository services for a country can offer significant advantages. In addition to the economy given by having a single point of technical support and development, standards are easier introduce and policy development may be simplified. One significant disadvantage is the possible depersonalization of the service, taking it away from local management and academic liaison.

METAFOR - Common Metadata for Climate Modelling Digital Repositories

The main objective of METAFOR [48] is to develop a Common Information Model (CIM) to describe climate data and the models that produce it in a standard way, and to ensure the wide adoption of the CIM. METAFOR will address the fragmentation and gaps in availability of metadata (data describing data) as well as duplication of information collection and problems of identifying, accessing or using climate data that are currently found in existing repositories.

METAFOR will optimize the way climate data infrastructures are used to store knowledge, thereby adding value to primary research data and information, and providing an essential asset for the numerous stakeholders actively engaged in climate change issues (policy, research, impacts, mitigation, private sector).

D4SCIENCE - Distributed Collaborative Infrastructure on Grid Enabled Technology for Science

D4Science [49] (DIstributed colLaboratories Infrastructure on Grid ENabled Technology 4 Science - Jan 2008-Dec 2009) is a project co-funded by European Commission's Seventh Framework Programme for Research and Technological Development involving 11 participating organizations. It aims at to continue the path that GÉANT [2], EGEE [3] and DILIGENT [4] projects have initiated towards establishing networked, grid-based, and data-centric e-Infrastructures that accelerate multidisciplinary research by overcoming barriers related to heterogeneity, sustainability and scalability.

In particular, D4Science is currently operating an infrastructure managing heterogeneous resources including hardware resources, i.e. machines acting as computing and storage resources (in part borrowed from the EGEE infrastructure); hosting environments, i.e. run-time containers supporting dynamic software deployment; software resources, i.e. software packages implementing specific functions; services, i.e. running instances of software resources providing functions; and data resources, i.e. collection of compound information objects representing various kinds of information. This infrastructure promotes the sharing of such variety of resources through the definition and dynamic creation of Virtual Research Environments (VRE). A VRE is an application consuming resources dynamically borrowed from the infrastructure, bound (and deployed) instantly, when and for the period they are needed.

This infrastructure is currently supporting the operation of two very large and challenging scientific communities: the Environmental Monitoring and the Fisheries and Aquaculture Resources Management communities. These scientific communities are served through three virtual organizations (VOs): Environmental Monitoring VO, Fishery Country Profiles Production System VO, and Integrated Capture Information System VO. These VOs are dynamic group of individuals and/or institutions defined around a set of sharing rules in which resource providers and consumers specify clearly what is shared, who is allowed to share, and the conditions under which sharing occurs to serve the needs of a specific community. These VOs consists of various resources including collection of Earth images, satellite products, species distribution maps, reports, statistical data, and tools for processing and analyzing them. They also support the definition and operation of various VREs.

The development and operation of the D4Science infrastructure is supported by the gCube software system. gCube is a distributed system for the operation of large-scale scientific infrastructures. It has been designed from the ground up to support the full life-cycle of modern scientific enquiry, with particular emphasis on application-level requirements of information and knowledge management. To this end, it interfaces pan-European Grid middleware (gLite) for shared access to high-end computational and storage resources, but complements it with a rich

array of services that collate, describe, annotate, merge, transform, index, search, and present information for a variety of multidisciplinary and international communities. Services, information, and machines are infrastructural resources that communities select, share, and consume in the scope of collaborative Virtual Research Environments.

Powered by gCube the D4Science infrastructure supports a very powerful and flexible notion of Information Object.

The gCube Information Object model resembles a very simple yet powerful graph-based data model whose constituents are gCube Information Objects (the nodes) and gCube Information Object Relationships (the edges).

Each gCube Information Object is characterised by:

- an identifier, i.e. a token attached to each object for identification purposes;
- a name, i.e. a word or set of words by which the object is known;
- a type, i.e. an attribute characterising the actual content of the object similarly to MIME type;
- a set of properties, i.e. a set of key-value pairs that can be used to attach additional attributes to the object;
- (optional) a raw-content, i.e. the actual information payload captured by the object.

Each gCube Information Object Relationship is characterised by:

- the source object, i.e. the object at which the relationship starts;
- the target object, i.e. the object at which the relationship ends;
- a primary role, i.e. the characterisation of the function played by the relationship;
- (optional) a secondary role, i.e. a specialisation of the function played by the relationship as expressed by the primary role;
- (optional) a position, i.e. an order-oriented attribute arranging relationships having the same source object according to a sequence;
- (optional) a set of properties, i.e. a set of key-value pairs that can be used to attach additional attributes to the relationship.

By properly instantiating this graph based model, gCube services support the notions of:

- Document, i.e. any gCube Information Object that should be considered as a primary unit information. It can be connected with other gCube Information Objects playing the role of its metadata, its annotation, its part and its alternative representations;
- Collection, i.e. a gCube Information Object source in a set of relationships having isMemberOf primary role. It can be further specialised in (i) Content Collection, in the case of Document target objects, (ii) Metadata Collection, in the case of Metadata target objects, and (iii) Annotation Collection, in the case of Annotation target objects;
- Annotation Object, i.e. a gCube Information Object target of a relationship having isDescribedBy primary role and isAnnotatedBy secondary role;
- Metadata Object, i.e. a gCube Information Object target of a relationship having isDescribedBy primary role.

GMES - Global Monitoring for Environment and Security

GMES [79] is the European participation in the worldwide monitoring and management of our planet Earth and the European contribution to the Group on Earth Observation (GEO). The global community acts together for a synergy of all techniques of observation, detection and analysis.

At the World Summit on Earth Observation in Washington in July 2003, the Group on Earth Observations (GEO) was established, with the goal of addressing the information requirement for the environment on a global scale. This work was completed in Brussels in February 2005 by the adoption of a 10 year implementation plan of an integrated Global Earth Observation System of Systems (GEOSS).

The GEOSS is an ambitious programme of information for ecological security and durable development intended for mankind. It principally foresees the monitoring and understanding of nature, the extent of disasters due to human activities, the impact of global warming, desertification, erosion and deforestation.

GMES will be the main European contribution to GEOSS.

The eSDI-NET+ project (Network for promotion of cross-border dialogue and exchange of best practices on Spatial Data Infrastructures throughout Europe) is co-funded by the European Community programme eContentplus (within DG Information Society and Media) for a period of 3 years (2007-2010). The project targets users and aims at gathering European Spatial Data Infrastructures (SDI) stakeholders in a platform for communication and knowledge exchange at all levels, with an emphasis on the user benefits. The purpose of this network is to raise awareness of the important role SDIs play and to promote cross border dialogue resulting in the creation of synthesised SDI guidelines and standards. By establishing communication mechanisms between European and local levels, this initiative will support better use of geographic information provided by European initiatives such as INSPIRE, GMES and GALILEO.

Recently, the eSDI-Net+ project has launched a broad assessment campaign consisting of identification and analysis of SDI's best practices at sub national level. This process will end up by the SDI Best Practice Award at the end of this year, during an international conference gathering the European communities involved in geo-information issues.

HMA - Heterogeneous Missions Accessibility

HMA [50] has the objective to define the interoperability concept across the ground segments of the European, Canadian and EUMETSAT missions which shall contribute to the initial phase of GMES. Focuses on interoperability of existing infrastructures and puts strong emphasis on metadata management (following OGC/ISO specifications and standards; the Open Geospatial Consortium, Inc.(OGC) is a non-profit, international, voluntary consensus standards organization that is leading the development of standards for geospatial and location based services.).

GEOLAND - Integrated GMES Project on Land Cover and Vegetation

GEOLAND [51] is carried out in the context of GMES, a joint initiative of European Commission (EC) and European Space Agency (ESA), which aims to build up a European capacity for Global Monitoring of Environment and Security.

The ambition of the geoland consortium is to develop and demonstrate a range of reliable, affordable and cost efficient European geo-information services, supporting the implementation of European directives and their national implementation, as well as European and International policies. Thus, the GMES initiative is considered a unique opportunity to integrate existing technology with innovative and scientifically sound elements into sustainable services.

Within eight sub-projects, the 56 geoland partners develop products and services, utilizing available Earth Observation resources in combination with in-situ measurements, and integrating them with existing models into pre-operational geo-information services. These will support international, European, national and regional authorities and institutions in fulfilling their increasing monitoring and reporting obligations - and help them to better manage natural resources.

MyOcean - Ocean Monitoring and Forecasting

MyOcean [52] is the implementation project of the GMES Marine Core Service, aiming at deploying the first concerted and integrated pan-European capacity for Ocean Monitoring and Forecasting. It provides the best information available on the Ocean for the large scale (worldwide coverage) and regional scales (European seas), based on the combination of space and in situ observations, and their assimilation into 3D simulation models. MyOcean wishes to offer the capability of combining and derive information from heterogeneous data types (space and in situ). The pan-European full fledged service will offer a single and reliable entry point to users and a direct access to products. The web portal will be directly connected to production units all over Europe to ensure homogeneity and full operationality. Service will include INSPIRE functionalities (discovering, visualization and downloading tools, ...) and a 24/7 HelpDesk. V1 Products will follow MyOcean Data Policy : open and free to any user and for any use.

INSEA - Data Integration System for Eutrophication

INSEA [53] is focused on the development of integrated management tools for coastal eutrophication assessment combining Models, Satellite Remote Sensing and in situ Measurements. INSEA aims to set-up and validate numerically robust ecological modelling systems in order to describe biogeochemical cycling of carbon and nutrients occurring under different hydrographical and trophic regimes, and to explore the system capabilities in a forecast mode to support coastal zone management issues. INSEA will provide innovation in the following areas:

- Assimilation techniques and sub-grid-scale processes knowledge: benefiting from large amounts of data made available by the project.
- Modelling & Data Managing Tools: Data acquisition, visualization and analysis.
- Management & Decision making: Forecasting capacity & better scenario evaluation and development of indexes.

One of the project deliverables, the "Metadata handbook" [78], provides metadata formats and data quality procedure guidelines to be followed by the INSEA partners.

MERSEA - Marine EnviRonment and Security for the European Area

MERSEA [54] aims to develop a European system for operational monitoring and forecasting on global and regional scales of the ocean physics, biogeochemistry and ecosystems. MERSEA provides for product information management: product catalogue and product standard ISO19115 description; product search facility.

PARSE.Insight - Permanent Access to the Records of Science in Europe

PARSE.Insight [55] is a two-year project co-funded by the European Union under the Seventh Framework Programme. It is concerned with the preservation of digital information in science, from primary data through analysis to the final publications resulting from the research. The problem is how to safeguard this valuable digital material over time, to ensure that it is accessible, usable and understandable in future. The rapid pace of change in information technology threatens media, file formats and software with obsolescence, and changing concepts and terminology also mean that, even if data can be read, it might not be correctly interpreted by future generations.

Many initiatives are already under way in this area, and the aim of the PARSE.Insight project is to develop a roadmap and recommendations for developing the e-infrastructure in order to maintain the long-term accessibility and usability of scientific digital information in Europe. The project will conduct surveys and in-depth case studies of different scientific disciplines and stakeholders and will base its results on these findings, as well as knowledge of ongoing developments.

PARSE.Insight is closely linked to the Alliance for Permanent Access to the Records of Science. The output from the project is intended to guide the European Commission's strategy about research infrastructure.

SOSI - Spatial Observation Services and Instrastructure

SOSI [56] is a project for developing innovative "Spatial Observation Services and Infrastructure" within the context of land monitoring initiatives at European and Member State levels. The project's objective is to demonstrate, in real operations, a decentralised information system allowing integration of distributed data and processing services as well as access and distribution at multiple levels, languages and content granularities.

The primary technology and operational procedures of SOSI are being implemented by utilizing the Service Support Environment (SSE) of ESA [2007]. SOSI offers a distributed node-based infrastructure of Web-services following Service Oriented Architecture (SOA) principles and standards thus establishing access to a number of content services and one land cover generation processing service operated by the participating organizations. The SSE infrastructure is providing coupling and user access mechanisms (binding, workflows and portal). The system includes enhanced user management and security features.

Climate-G

Climate-G [57] is a distributed testbed for climate change addressing challenging data and metadata management issues at a very large scale. The main scope of Climate-G is to allow scientists to carry out geographical and cross-institutional data discovery, access, visualization and sharing of climate data.

The Climate-G testbed is the result of an open, successful and wide collaboration joining grids and P2P paradigm, OGC services, visualization tools, etc. To enable geographical data sharing, search and discovery activities (through the Climate-G data grid portal interface) we adopted a distributed CMCC metadata solution leveraging P2P and grid technologies, the GRelC Data Access and Integration Service.

The Climate-G testbed provides a proof of concept concerning the involved technologies and right now it manages about 2TB of data provided by IPSL and University of Cantabria. Other datasets come from the IPCC website (AR4). Additional data from CMCC will be soon added to the digital library.

The Climate-G Data Distribution Centre is the data grid portal of the testbed and it is intended for scientists and researchers that want to carry out search and discovery activities on the available large scale digital library. It provides a ubiquitous and pervasive way to ease data publishing, metadata search & discovery, metadata annotation and validation, data access, etc.

The Climate-G data portal security model includes the use of HTTPS protocol for secure communication with the client (based on X509v3 certificates that must be loaded into the browser), secure cookies to establish and maintain user sessions as well as a complete role-based authorization system.

DEGREE - Dissemination and Exploitation of GRids in Earth science

DEGREE [58] aims to build a bridge linking the Earth Science (ES) and GRID communities throughout Europe. An ES applications panel with a range of candidate applications suitable for porting to GRID will make sure key ES requirements for porting and deployment on the GRID middleware are identified, communicated and discussed within the GRID community. Work Package 2 is dealing with data management and will perform a review of mostly used tools (database products), systems and standards for ES metadata in the ES community and will also evaluate and review advanced grid technology components to securely distribute, share and access ES data in an operation oriented environment.

In order to ensure that ES requirements are taken into account in the next Grid generation, DEGREE will initiate different collaborations; at short, medium and long term via EU horizontal collaborations, specific collaboration with Grid projects and participation to the e Infrastructure Reflection Group (e-IRG).

SeaDataNet

Pan-European [59] infrastructure for Ocean and Marine Data Management, provides integrated on-line access to a very comprehensive sets of multidisciplinary in-situ and remote sensing marine data, meta-data and products. Particular attention to interoperability, achieved by adopting the ISO 19115 metadata standard and using common vocabularies, harmonised Data Transport Formats for data sets delivery and SOAP Web Services.

SeaDataNet is an infrastructure, which interconnects the National Oceanographic Data Centres and marine data focal points from 35 countries around European seas. These data centres are part of major and leading marine research institutes in these countries and have a long-term perspective. This implicates that the SeaDataNet infrastructure also has a long-term perspective as an operational and well embedded network of centres and systems. The SeaDataNet project has a five year funding from the EU and is used to develop common protocols, standards and software modules and to develop and populate joint metadatabases and data access portals.

SeaDataNet is a multi-year activity. In terms of OGC standards, Version 1: from February 2008 onwards: Users can search and browse in common metadatabases including geographical interface (OGC compliant) and then request access to data by downloading services via a common shopping mechanism annex download manager. The central tool will arrange that requests are distributed to the data centres and that the data sets are provided for downloading by a common interface and in common formats. The metadatabases will also be available as Web services for feeding local and regional portals. All metadata are compliant to the ISO 19115 metadata standard. The ODV package has a seamless connection to SeaDataNet output formats.

In the coming years the SeaDataNet infrastructure will be extended into Version 2 with a further development of:

- Viewing services = Quick views and Visualisation of data and data products
- Product services = Generic and standard products

For the viewing services the OGC standards will be adopted, comprising Web Map Services (WMS), Web Feature Services (WFS) and Web Coverage Services (WCS). These will support the quick viewing and visualization of data sets and data products.

EMODNET - European Marine Observation and Data Network

EMODNET [60] aims to facilitate access to coherent data sets, to permit the recognition of data gaps and to shape a data collection and monitoring infrastructure directly suited to multiple applications. EMODNET is a European Commission initiative (DG MARE) to improve Europe's marine distributed data infrastructure launched in the Commission's Green Paper on maritime policy (COM(2006) 275 final). The project is developing standards across disciplines as well as within them.

Data on oceans and seas are available from many sources but assembling them for particular applications takes considerable effort and there is no overall policy for keeping them for posterity. An objective of the EU's new maritime policy is to integrate existing, but fragmented initiatives in order to facilitate access to primary data for public authorities, maritime services, related industries and researchers. The Commission has therefore undertaken to set up a European Marine Observation and Data Network to open up opportunities for high technology commercial companies in the maritime sector, improve the efficiency of activities such as marine observation, management of marine resources and marine research in European laboratories.

It will be assisted by a Marine Observation and Data Expert Group (MODEG), whose mission is to provide the Commission with the scientific, technical and operational expertise it needs to ensure that the European Marine Observation and Data Network (EMODNET) best meets the needs of its future users. The tasks of the Expert Group will be set by the Commission and will evolve. Initially this includes:

- assisting in the monitoring of all phases of studies, pilot projects and preparatory actions set up to further the aims of the European Marine Observation and Data Network.
- developing an agreed description of marine data that might be included within the Network.

HIDDRA - Highly Independent Data Distribution and Retrieval Architecture

Institutions such as NASA, ESA or JAXA find solutions to distribute data from their missions to the scientific community, and their long term archives. This is a complex problem, as it includes a vast amount of data, several geographically distributed archives, heterogeneous architectures with heterogeneous networks, and users spread

around the world. HIDDRA [62] proposes a novel architecture that solves this problem aiming to fulfil the requirements of the final user. The architecture is a modular system that provides a highly efficient parallel multi-protocol download engine, using a publisher/subscriber policy which helps the final user to obtain data of interest transparently.

GENISI-DR - Ground European Network for Earth Science Interoperations

GENISI-DR [63] has the challenge of establishing open Earth Science Digital Repository access for European and world-wide science users.

GENESI-DR shall operate, validate and optimise the integrated access and use available digital data repositories to demonstrate how Europe can best respond to the emerging global needs relating to the state of the Earth, a demand that is unsatisfied so far.

GENESI-DR has identified the following objectives:

- To provide guaranteed, reliable, easy, effective, and operational access to a variety of data sources, and demonstrate how the same approach can be extended to provide access to all Earth Science data
- To harmonise operations at key Earth Science data repositories limiting fragmentation of solution
- To demonstrate effective curation and prepare the frame for approaching long term preservation of Earth Science data
- To validate the effective capabilities required to access distributed repositories for new communities, including education, and assess benefits and impacts
- To integrate new scientific and technological derived paradigms in operational infrastructures in responds to the latest Earth Science requirements

GENESI-DR builds upon the existing, operational and focused Earth Observation (EO) European infrastructure and involves key Earth Science centres responsible for operational data acquisition, processing, archiving and distribution.

The implementation and the systematic monitoring of international Environmental conventions need data, tools and world-wide infrastructures to gather and share the data.

A common dedicated infrastructure will permit the Earth Science communities to derive objective information and to share knowledge in all environmental sensitive domains over a continuum of time (from historical measurement to real time assessment to short and long term predictions) and a variety of geographical scales (from global scale to very local facts).

Furthermore, each specific Earth Science domains community has existing methods, approaches and working practices for gathering, storing and exchanging data and information. These are likely to impose a considerable constraint on the impact and increased effectiveness generated by a shared e-Infrastructure approach.

The challenge in front of us today, is to offer a framework that allows scientists from different Earth Science disciplines to have access, to combine and to integrate all historical and present Earth-related data from space, airborne and in situ sensors available from all digital repositories dispersed all over Europe together.

LIFEWATCH

The research infrastructure LifeWatch [65] will construct and bring into operation the facilities, hardware, software and governance structures for all aspects of biodiversity research. It will consist of: facilities for data generation and processing; a network of observatories; facilities for data integration and interoperability; virtual laboratories offering a range of analytical and modelling tools; and a Service Centre providing special services for scientific and policy users, including training and research opportunities for young scientists. The infrastructure has the support of all major European biodiversity research networks.

Rationale While we are exploring other planets, its is surprising how little we still know about our own planet Earth. This is especially true for our understanding of the living world, the biological diversity of species, and their genes and the ecosystems in which they occur. We only know a fraction of the probably millions of species, especially of the insects, microorganisms and other small species which are in different ways crucial for goods and services such as pollination, health or biotechnology. Scientific developments generated knowledge about some components of biodiversity, but the research community needs a new methodological approach to understand the biodiversity system. LifeWatch is designed to serve science as a large-scale facility offering on-line facilities for such methodologies.

Vision and ambition The LifeWatch infrastructure for biodiversity research addresses the huge gaps we face in our understanding of life on Earth. Its innovative design supports a large-scale methodological approach to data resources, advanced algorithms and computational capability. Life Watch will not only serve to support the scientific research, but also support the understanding and the rational management of our ecosystems by policy makers, the private sector and the general public.

Mission and services

- Operate a single portal for pure and applied researchers, policy makers, industries and the public at large
- Enable new scientific practices and inspire a new generation of scientists
- Structure the scientific community with new opportunities for large scale projects and data capture priorities
- Offer knowledge-based decision-support for the rational management of our ecosystems on land and in the seas, and to policy makers and the public
- Innovate biodiversity based industry towards sustainable practices
- Educate to catalyze the discovery and innovation process
- Provide on-line and off-line user support

Architecture The wealth of large data sets on the different levels of biodiversity (genes, populations, species and ecosystems) opens up an unprecedented new area of research. But the complex and multidisciplinary problems also force scientists to collaborate in virtual organizations at a global scale. LifeWatch will enable 'distributed large-scale' science, which is the only way to participate in new developments in biodiversity science.

User groups can create their own e-laboratories or e-services within the common architecture of the infrastructure. They may share their data and analytical and modeling algorithms with others, while controlling access. All public resources, such as data repositories, computational capabilities and capacity are available through the problem solving environment.

The architecture allows for dynamic linkages to other resources and associated infrastructures. As such, LifeWatch is an example of the new generation of research infrastructures that form a cooperating fabric.

International cooperation The implementation of LifeWatch is only possible through international cooperation. The sheer size of the infrastructure with respect to costs, functionalities and user communities requires large-scale collaboration. The European Strategy Forum on Research Infrastructures (ESFRI) identified LifeWatch as an essential facility to be supported by European countries. Currently 19 European countries expressed their interest in LifeWatch.

LifeWatch cooperates with other international infrastructures to add value to its services. Such cooperation includes the connection to a variety of data repositories, or providing analysis of ground-level data with Earth-observation data from satellites. At the global level the LifeWatch partners interact with the relevant knowledge centres to work together within a cost-effective and targeted long-term strategy in order to serve the global interests.

Issues

Data

- LifeWatch combines many small-to-large datasets primarily containing taxonomic and ecological data, but also including climatological, meteorological, topographic, soil composition, land use, human factors and other datasets.
- much of the data has geospatial attributes, which may be in different Coordinate Reference Systems. Performing queries over this data is complex.
- similarly, taxonomic data can conform to multiple models, making queries complex, and in many cases, ambiguous.

Metadata

- since there is not necessarily edit access to the primary datasets, metadata must be stored separately to the data. It is expected that there may be multiple repositories of different metadata about the same data. It has to be ensured that metadata is linked to the data, and if possible, data linked to metadata, and that queries can be performed efficiently.

Infrastructure

- a list of candidate unique identification schemes has been identified that could be used to identify objects within the LifeWatch system. Although they have slightly different properties, there is no compelling reason to select a particular scheme. Agreement by other projects and infrastructure support might provide a compelling reason.
- several authentication, authorization and accounting requirements, which are partially supported in current infrastructures, need further development.

BioCASE

The Biological Collection Access Service for Europe[66] , is a transnational network of biological collections of all kinds. BioCASE enables widespread unified access to distributed and heterogeneous European collection and observational databases using open-source, system-independent software and open data standards and protocols.

During BioCASE‘ EU-funded project phase (2001-2004), partners from 31 countries established the network, starting with meta-information on thousands of biological collections, and followed by a unit-level access network. “Unit-level” data refer to individual collection or observation units, i. e. individual specimens or observation records. In contrast, “collection-level metadata” consist of records describing entire collections of such units. This information comes in formats ranging from XML and text data to high-resolution images and even video files.

The continuous development of BioCASE, as well as user and data provider support, was or is supported by the European Union projects ENBI and SYNTHESYS, as well as by other initiatives such as the GBIF mirror and replication project.

GBIF - Global Biodiversity Information Facility

GBIF [67] is an international organisation that is working to make the world’s biodiversity data accessible everywhere in the world. GBIF and its many partners work to mobilise the data, and to improve search mechanisms, data and metadata standards, web services, and the other components of an Internet-based information infrastructure for biodiversity.

One of GBIF’s main purposes is enabling a global distributed network of interoperable databases that contain primary biodiversity data. By this we mean data associated with specimens in biological collections, as well as documented observations of plants and animals in nature. Such data nearly always have certain common attributes, such as scientific name, location, and date of collection. Therefore, they can easily be recast into a common data exchange format, and shared with the world using internet protocols that recognize and handle these data exchange formats.

GBIF recommends the adoption of either Darwin Core 1.4 (also known as Darwin Core 2) with the DiGIR protocol package, or the ABCD and the BioCASe protocol package.

EDIT - European Distributed Institute of Taxonomy

EDIT [68] is the collective initiative of 27 leading European, North American and Russian institutions to reduce the fragmentation in taxonomic research through integration of capacities and activities in a joint Network of Excellence.

Among the members of EDIT are the premier natural history collections-based institutions worldwide, which have both the management capacity and the will to progress toward EDIT's objectives. Their collections are global in coverage and are supported by complementary expertise. More than half of the world's natural history specimens, which constitute the large scale infrastructure for taxonomic research, are held in the repositories of EDIT's membership. The inclusion in the consortium of network institutions devoted to management of biodiversity data, and of a research organisation directly related to users of taxonomy for agriculture and environment, in addition to the links or inclusion of many partners with universities will facilitate dissemination of EDIT taxonomic research and training toward a wide audience.

EDIT will bring together the leading taxonomic institutions in Europe that for historical reasons have developed independently. The association with leading North American and Russian partners will make it a worldwide leading network. The consortium so constituted unites the premier natural history collections-based institutions, to progress toward EDIT's structural and scientific objectives.

ENBI - European Network for Biodiversity Information

ENBI [69] is a network pooling the technical resources and human expertise in biodiversity informatics within Europe. ENBI enhances the communication and co-operation between GBIF-nodes, biodiversity institutes and relevant initiatives in Europe.

ENBI is the European contribution to the Global Biodiversity Information Facility (GBIF). The business plan of GBIF gives priority to the vast objective to make primary biodiversity data globally available. In first instance GBIF is restricted to taxonomic data and to biological collection and specimen data, as well as to promoting the common access and interoperability between these databases. ENBI follows these priorities by concentrating on databases at the European scale and on activities that need co-operation at a European level. ENBI also explores the potential of tools to apply the biodiversity data as such, or in combination with other categories of data. In addition, ENBI focuses on the market of end-users with special attention on processes to develop specific products and services. A strong network will be established and maintained by bringing together the major stakeholders in the field of biodiversity information. Members of the network are the co-ordinating institutes of past and current EU biodiversity projects, and the (designated) institutes that act as, or host, the national GBIF-nodes. Also the NAS countries are represented in ENBI. The activities of ENBI are co-ordinated with those of the European Community Clearing-House Mechanism and the European Environmental Agency.

MARBEF - Marine Biodiversity and Ecosystem Functioning

MarBEF [71] is a EU Network of Excellence consisting of 83 European marine institutes The platform integrates and disseminates knowledge and expertise on marine biodiversity, with links to researchers, industry, stakeholders and the general public.

Information on the existence of data is a prerequisite to data sharing. MarBEF will inventory all aspects of marine science relevant to marine biodiversity; in this collaboration will be sought with existing initiatives such as MEDI and EDMED. A database will be created storing information on planned, ongoing and finalised research. Lists of institutions, scientists and their expertise and publications will be maintained. A database of research facilities, oceanographic vessels and cruises will allow efficient sharing of resources. The most important type of information will consist of an inventory of existing biodiversity databases. These will be documented, giving details on taxonomic scope, information content, access constraints and quality, and where possible a direct link provided to an internet entry to the data.

A second step covers the need to create massive data sets on species and biogeography. MarBEF will establish a public European warehouse of biological databases together with intelligent tools for data mining. Development will be undertaken so that marine biological data sets will be accessed at their host institutes by the central MarBEF portal to build maps showing distribution of species and biotopes. The inventory of biodiversity data sets will be instrumental in identifying candidate data sets for integration. The work will be undertaken with and take advantage of development already undertaken by the Ocean Biodiversity Information System (OBIS) so that a facility will be developed known as EurOBIS.

To facilitate exchange of research results, data formats and standards will be proposed; in this, efforts of other groups will be re-utilised. Links will be established with TDWG, ICES/MDM, IODE/GE-TADE, IODE/GE-BCDMEP and others for data standards for biological data. The taxonomic backbone to all the data activities will be the European Register of Marine Species. Mechanisms will be developed to keep the ERMS up-to-date, and synchronous with major data contributors such as ETI's World Biodiversity database, FishBase, Algaebase, CLEMAM and many others. ERMS will be available as a fully searchable database though two or more web sites. The use of a central register of taxonomic names will improve the quality of integration of data from several sources. MarBEF will represent its members in international activities on data management, such as IOC, OBIS and GBIF; based on its accumulated data and tools, it will be able to play a leading role in such initiatives.

Marine Genomics Europe

The EU Network of Excellence Marine Genomics [72] was established for the implementation of high-throughput genomic approaches to support the biology of marine organisms. The network unites 44 institutions from 16 countries (within and outside Europe) with 450 participating scientists.

The Marine Genomics Europe network will bring high-throughput approaches to the study of marine organisms. Within the platform, the MGE Bioinformatics Portal provides a central point of reference for all data sets and tools. It provides data storage, data analysis and data integration. The Marine Genomics infrastructure was developed as a clearing house of functional genomics data for marine organisms. It currently includes tools to upload, preprocess, cross-reference, annotate, NCBI submit and store EST data. It also includes a corresponding microarray design, upload and storage tool with development of analysis tools underway.

The usage of the Marine Genomics infrastructure has been speedily increasing with more species databases being added and the numbers of sequences increasing even faster. Furthermore, Marine Genomics integrates the microarray entries with the MATLAB environment for which there are several commercial and public libraries ("toolboxes") for microarray analysis.

Current development goals include the added functionality of continued addition of ontology and information for all species. Another goal is to add the ability to parse and process multiple microarray platforms such that users can have the flexibility of uploading data output from their own individual microarray platforms. Finally, in particular special care will be given to speedily exporting expression data as MAGE-XML for incorporation in NCBI's GEO databases rather than having that data exclusively retained in Marine Genomics.

The ultimate purpose of Marine Genomics is indeed to assist in submitting quality data to the NCBI GenBank and GEO databases. For that purpose, the Marine Genomics pipeline and tools have been assembled to provide a medium for working with functional genomics in a marine biology environment.[75]

SYNTHESYS - Synthesis of systematic resources

SYNTHESYS [73] is a EU- funded initiative comprising of 20 European natural history museums and botanic gardens, and aims to create an integrated infrastructure for researchers in the natural sciences.

SYNTHESYS aims to raise scientists' awareness of best practice in handling and sampling collections by offering improved training and workshop opportunities, and guidelines for the care, storage and conservation of collections. It will create an integrated European resource; bringing together the biological and geological collections held by major natural history museums and other institutions.

APA - Alliance of Permanent Access

The Alliance of Permanent Access [76] aims to develop a shared vision and framework for a sustainable organisational infrastructure for permanent access to scientific information.

The Alliance advocates a pragmatic approach both to the idea of an infrastructure that serves all scientific fields and to the way to accomplish this. The basic notion is the assumption that the repository infrastructure will be different for each scientific "community", such as particle physics, astronomy and space science, life sciences, earth and environmental sciences, or social sciences. The repositories will be part of organisations that exist within a particular community. But all these communities must agree on certain standards to make their repositories interoperable. The repositories will also benefit from a number of common facilities: from common R&D activities or a framework that offers technical tools, to a single accreditation body for guaranteeing the quality of companies or organisations that verify whether a repository meets the requirements. The repositories themselves are physical, but the overall infrastructure is virtual: there is no need for a central governing body, though some kind of competence centre to support the common goals makes sense. The key to the approach of the Alliance is the engagement with the various scientific "communities" to help them build their part of the infrastructure.

The practical experience of the Alliance members is very important here. The Alliance can also establish the connections to a variety of parties, including policy makers and funding bodies. This way, communities will be assisted in some crucial preservation activities, like:

- identifying which digital information to preserve and for how long;
- identifying key repositories;
- establishing a scheme for metadata suitable for the community;
- organising test beds, etc.

The pragmatic approach of the Alliance is also shown by the choice to start with three or four communities that are already reasonably well-organised. The Alliance believes that this will generate sufficient momentum for others to follow. And in developing common tools for this small number of initial communities, it is easy to check for their wider suitability.

DELOS

DELOS [77] is a Network of Excellence on Digital Libraries partially funded by the European Commission in the frame of the Information Society Technologies Programme (IST). The main objectives of DELOS are research, whose results are in the public domain, and technology transfer, through cooperation agreements with interested parties.

DELOS is currently working on the development of a Digital Library Reference Model that is designed to meet the need of the next-generation systems, and on a globally integrated prototype implementation of a Digital Library Management System, called Delos DLMS, which will serve a s a concrete partial implementation of the reference model and will encompass many software components developed by DELOS partners

PDB - Protein Data Bank

The PDB archive [82] contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the wwPDB, the RCSB PDB curates and annotates PDB data according to agreed upon standards.

The PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

The PDB archive is the single worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids. These are the molecules of life that are found in all organisms including bacteria, yeast, plants, flies, other animals, and humans. Understanding the shape of a molecule helps to understand how it works. This knowledge can be used to help deduce a structure's role in human health and

disease, and in drug development. The structures in the archive range from tiny proteins and bits of DNA to complex molecular machines like the ribosome.

The PDB was established in 1971 at Brookhaven National Laboratory and originally contained 7 structures. In 1998, the Research Collaboratory for Structural Bioinformatics (RCSB) became responsible for the management of the PDB. In 2003, the wwPDB was formed to maintain a single PDB archive of macromolecular structural data that is freely and publicly available to the global community. It consists of organizations that act as deposition, data processing and distribution centres for PDB data.

In addition, the RCSB PDB supports a website where visitors can perform simple and complex queries on the data, analyze, and visualize the results. Details about the history, function, progress, and future goals of the RCSB PDB can be found in our Annual Reports and Newsletters.

The PDB Advisory Notice defines the conditions for using data from the PDB archive. Our Policies & References page describes copyright restrictions on RCSB PDB materials, our privacy policy, and citation information. Data deposition and release policies are available from deposit.pdb.org.

The RCSB PDB has an international community of users, including biologists (in fields such as structural biology, biochemistry, genetics, pharmacology); other scientists (in fields such as bioinformatics, software developers for data analysis and visualization); students and educators (all levels); media writers, illustrators, textbook authors; and the general public.

The RCSB PDB website at www.pdb.org is accessed by about 140,000 unique visitors per month from nearly 140 different countries. Around 500 GigaBytes of data are transferred each month. Data are accessed via the website, ftp server (supporting ftp and rsync access), Web Services and RSS feeds.

EuroVO-AIDA - European Virtual Observatory-Astronomical Infrastructure for Data Access

EuroVO-AIDA [83] aims at unifying the digital data collections of European astronomy, integrating their access mechanisms with evolving e-technologies, and enhancing the science extracted from these datasets. The EuroVO-AIDA project is proposed to lead the transition of Euro-VO into an operational phase.

EuroVO-AIDA integrates the technology, networking and service activities of Euro-VO into a fully functioning structure. It will establish a Registry of VObs-compliant resources; support the network of Data Centres in deploying the VObs eInfrastructure; co-ordinate development of user tools for science extraction; and disseminate results to the astronomical community and identify their needs. The VObs interoperability standards will be updated taking into account feedback from implementation by data centres and from science usage. Specific emphasis will be placed on data access and data models, and on assessing innovative use of emerging technologies such as web 2.0 by data centres for continuous improvement of the eInfrastructure.

The result will be an operating knowledge infrastructure that enables and stimulates new scientific usage of astronomy digital repositories.

The Virtual Observatory aims to provide the framework for global access to the various data archives by facilitating the standardisation of archiving and data-mining protocols. The AVO will also take advantage of state-of-the-art advances in data-handling software in astronomy and in other fields.

The Virtual Observatory initiative is currently aiming at a global collaboration of the astronomical communities in Europe, North and South America, Asia, and Australia under the auspices of the recently formed International Virtual Observatory Alliance.

HEP - High Energy Physics

HEP is, and has always been, a major data producer. A first workshop on data preservation and long-term analysis in High-Energy Physics was held at DESY, Hamburg, on 26-28 January 2008 with about 50 physicists and IT experts participating. A short report of the workshop can be found at [84].

Please find hereafter an extract:

High-energy physics (HEP) experiments acquire huge datasets that may not be superseded by new and better measurements for decades or centuries. Nevertheless, the cost and difficulty of preserving both the data and the understanding of how to use them are daunting. The small number of cases in which data over ten years old have been reanalysed has only served to underline that such analyses are currently very close to impossible. The recent termination of data taking by the H1 and ZEUS experiments at DESY's HERA collider, and by the BaBar

experiment at SLAC, plus the imminent termination of other major experiments, prompted the organisation of this workshop.

The workshop heard from HEP experiments long past ('it's hopeless to try now'), recent or almost past ('we really must do something') and included representatives form experiments just starting ('interesting issue, but we're really very busy right now'). We were told how luck and industry had succeeded in obtaining new results from 20-year-old data from the JADE experiment, and how the astronomy community apparently shames HEP by taking a formalised approach to preserving data in an intelligible format. Technical issues including preserving the bits and preserving the ability to run ancient software on long-dead operating systems were also addressed. The final input to the workshop was a somewhat asymmetric picture of the funding agency interests from the two sides of the Atlantic.

Parallel working sessions addressed the different needs and issues in e+e-, ep and pp experiments. The likelihood of future data rendering old datasets uninteresting can be very different in these three types of collision. The workshop then tried to lay out a path forward. Apart from the obvious 'hold another workshop' imperative, it seemed clear that experimental programmes that felt 'we really must do something' should be used as drivers. A first step must be to examine issues of cost, difficulty and benefit as quantitatively as possible so that the next workshop could have concrete discussions on the science case for various levels of data preservation.

ICOS - Integrated Carbon Observation System

ICOS [89] is a new strategic research infrastructure to quantify the greenhouse gas balance of Europe and adjacent regions. It consists of a harmonized network of ecosystem long-term observation sites, a network of atmospheric greenhouse gas concentration sites and a network of ocean observations. The networks will be coordinated through a set of central facilities, including an atmospheric, an ecosystem and an oceanic thematic centre, a central data centre (the ICOS portal), and an analytical laboratory. ICOS will provide the essential long-term observations required to understand the present state and predict future behaviour of the global carbon cycle and greenhouse gas emissions. It will monitor and assess the effectiveness of carbon sequestration and/or greenhouse gases emission reduction activities on global atmospheric composition levels, including attribution of sources and sinks by region and sector.

International landscape ICOS will contribute to the implementation of the Integrated Global Carbon Observation System (IGCO). At the same time, ICOS fulfils the monitoring obligations of Europe under the United Nations Framework Convention on Climate Change (UNFCCC). The list of variables covered in ICOS are central to GEOSS (Global Earth Observation System of Systems) as recommended to 'support the development of observational capabilities for Essential Climate Variables (ECVs)'. Further, ICOS contributes to the GEOSS aims by implementing in Europe the IGOS-P (Integrated Global Observing Strategy - Partnership) for Atmospheric Chemistry Observations (IGACO) and for Integrated Global Carbon Observations (IGCO). ICOS will contribute to the European share of global greenhouse gas observations under GEO (Group on Earth Observations), WMO-GAW (World Meteorological Organisation-Global Atmosphere Watch), and GTOS (Global Terrestrial Observing System) programs.

Data Management ICOS data is centralised in the atmospheric, ecosystem and oceanic thematic centres. It can be accessed through a single entry point called the ICOS carbon portal.

Data Product levels definitions

- Level 0: Raw data (e.g. current, voltages) produced by each measuring instrument
- Level 1: Data expressed in geophysical units. Two distinct data streams will be generated. A Rapid Delivery Data (RDD) stream and a Long Term Data (LTD) stream with higher precision.
- Level 2: Elaborated time series (e.g. gap-filling, combined with meteorological, local flux and Boundary Layer Height observations) and other products (e.g. PBL, meteo data ...)

Databases The databases will handle input of data from the station networks of about 60 to 100 main sites, possibly of hundreds of additional sites when the research infrastructure will be fully operational during the next 20 years. Besides measuring instrument data, the database will store site metadata, calibration tanks information (including logistical information allowing to track their circulation between the network stations and the analytical laboratory). Database software modules will be developed to provide site metadata directly from the ICOS atmospheric stations. The database archiving will be dynamic, complete and robust over time to allow for an automatic reprocessing of the whole dataset for instance when primary scale changes will occur (every 3 years on average). Off site backup will be properly dimensioned and operated. A web site, interfacing with the database, will provide a fast, dynamic and robust way to handle user data requests. The database model will be able to track the changes in the data and in the processing programs using versioning. ICOS will establish and lead strong relations with other databases containing similar data in other continents to enhance cross network standardization.

Accessibility ICOS will implement a data policy in accord with the INSPIRE directive. The general thrust of the INSPIRE directive is towards timely, usable and open access to spatial and other data within Europe while protecting the rights of data owners. It also requires that data comply with standards to ensure easy data discovery and sharing. Most of these are met naturally by the proper structuring of the ICOS database using open software standards.

4 DMTF-SURVEY MEMBERSHIP

First Name	Name	Country	eMail	Project-Expertise
Patrick	Aerts	NL	aerts@nwo.nl	HPC
Andreas	Aschenbrenner	DE	aschenbrenner@sub.uni-goettingen.de	DARIAH
Lalos	Balint	HU	lajos.balint@niif.hu	NREN
Hilary	Beedham	UK	beedh@essex.ac.uk	CESSDA
Victor	Castelo	ES	victor.castelo@cti.csic.es	CSIC
Brian	Coghlan	IE	coghlan@cs.tcd.ie	HPC
Rudolf	Dimper	EIRO	dimper@esrf.fr	ESRF
Luigi	Fusco	EIRO	luigi.fusco@esa.int	ESA
Francoise	Genova	FR	genova@newb6.u-strasbg.fr	CDS
David	Giaretta	UK	david.giaretta@stfc.ac.uk	CASPAR, PARSE.Insight
Jonathan	Giddy	UK	j.p.giddy@wesc.ac.uk	LIFEWATCH
Maria	Koutrokoi	GR	mkoutr@gsrt.gr	NCP-RI
Carlos	Morais-Pires	EC	carlos.morais-pires@ec.europa.eu	EC
Christian	Ohmann	DE	Christian.Ohmann@uni-duesseldorf.de	ECRIN
Pasquale	Pagano	IT	pasquale.pagano@isti.cnr.it	D4Science
Leonard	Rivier	FR	leonard.rivier@isce.ipsl.fr	ICOS
Lorenza	Saracco	EC	lorenza.saracco@ec.europa.eu	EC
Dany	Vandromme	FR	dany.vandromme@renater.fr	RENATER/ESFRI
Peter	Wittenburg	NL	peter.wittenburg@mpi.nl	CLARIN/ESFRI
Hans	Zandbelt	NL	hans.zandbelt@surfnet.nl	SURFNET
Matti	Heikkurinen	CN	matti@emergence-tech.co.uk	DMTF Support
Michèle	Landes	FR	landes@renater.fr	DMTF Support
Ana Bela	Sa Dias	NL	dias@nwo.nl	DMTF Support

Table 1.1: DMTF-SURVEY Membership

5 Definitions, Acronyms and Abbreviations

DCMI Dublin Core Metadata Initiative, an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. http://dublincore.org/

DDI Data Documentation Initiative. http://www.ddialliance.org/

DMTF-INTEROP e-IRG Data Management Task Force Interoperability Subgroup

DMTF-SURVEY e-IRG Data Management Task Force Survey Subgroup

DOI Digital Object Identifier, a persistent identifier for a document that can be handled by a resolution service to direct communication to the correct server. Developed by the International DOI Foundation.

DRM Digital Rights Management

DSA Data Seal of Approval (see Appendix A)

e-IRG e-Infrastructures Reflection Group

EHR Electronic Health Record

EMR Electronic Medical Record

ESFRI European Strategy Forum on Research Infrastructures

EU European Union

GIS Geographical Information System

JISC Joint Information Systems Committee. JISC supports further and higher education by providing strategic guidance, advice and opportunities to use ICT to support teaching, learning, research and administration. http://www.jisc.ac.uk/

OA Open Access

OAI Open Archives Initiative, develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content.

OAI-DC an XML format for the serialisation of Simple Dublin Core metadata descriptions. The format is defined as a "metadata format" for use within the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). OAI-PMH requires that data providers support the oai_dc metadata format.

OAI-PMH Open Archives Initiative Protocol for Metadata Harvesting; widely used, if not de facto standard protocol for harvesting metadata from OA repositories. Website includes a tutorial. http://www.oaforum.org/

OAIS Open Archival Information System reference model, a conceptual framework for an archival system dedicated to preserving and maintaining access to digital information over the long term.

OpenDOAR Open Directory of Open Access Repositories, Joint project of the Universities of Nottingham and Lund to create an authoritative reference database of Open Access repositories worldwide. http://www.opendoar.org/

OpenURL A standard for linking URLs that can be processed by link resolvers to direct users to the most appropriate copy of a resource. Originally created by Ex-Libris, http://www.exlibrisgroup.com/sfx_openurl.htm/

OSI Open Source Initiative, a non-profit corporation dedicated to managing and promoting the OSI Certified Open Source Software. http://www.opensource.org/

SOAP Simple Object Access Protocol, a web services-based protocol for querying Internet indexes or databases and returning search results

ToR Terms of Reference

6 REFERENCES

[1] ESFRI, European Roadmap for Research Infrastructures, Roadmap 2008, http://cordis.europa.eu/esfri

[2] M. Foulonneau et al: Digital Repositories Infrastructure Vision for European Research - Review of technical Standards, Driver Project, http://www.driver-support.eu/documents/DRIVER_Review_of_Technical_Standards.pdf

[3] Digital Repository Infrastructure Vision for European Research, http://en.wikipedia.org/wiki/Digital_Repository_Infrastructure_V

[4] Developing the UK's e-infrastructure for science and innovation - Report of the OSI e-Infrastructure Working Group, http://www.nesc.ac.ui/documents/OSI/report.pdf

[5] Digital Preservation Policies Studies, Charles Beagrie Ltd, A study funded by JISC http://www.jisc.ac.uk/media/documents/programmes/preservation/jiscpolicy_p1finalreport.pdf

[6] UKDA - UK Data Archive, Managing and Sharing Data - A best practice guide for researchers, http://www.data-archive.ac.uk/news/publications/managingsharing.pdf

[7] Francine Berman: Got data? A guide to data preservation in the information age. Communications of the ACM, December 2008, http://portal.acm.org/citation.cfm?doid=1409360.1409376

[8] Harnessing the Power of Digital Data for Science and Society. Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. January 2009, http://www.nitrd.gov/about/Harnessing_Power_Web.pdf

[9] Paul N. Edwards et al (eds): Understanding Infrastructure: Dynamics, Tensions, and Design. Report of a Workshop on "History + Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures", January 2007, http://deepblue.lib.umich.edu/bitstream/2027.42/49353/3/UnderstandingInfrastructure2007.pdf

[10] Tom Wilson, Institutional open archives, where are we now? Chartered Institutes of Library and Information Professionals, http://www.cilip.org.uk/publications/updatemagazine/archive/archive2006/april/tomWilsonApril06.htm

[11] Soo Young Rieh et al: Census of Institutional Repositories in the US. A comparison across institutions at different stages of IR development, http://www.dlib.org/dlib/november07/rieh/11rieh.html

[12] Clifford A. Lynch et al: Institutional Repository Deployment in the United States as of Early 2005, http://www.dlib.org/dlib/september05/lynch/09lynch.html

[13] Mary Anne Kennan, Danny A. Kingsley: A snapshot of Australian institutional repositories, http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2282/2092

[14] Craig Lee, George Percivall: Standards-Based Computing Capabilities for Distributed Geospatial Applications, http://www2.computer.org/portal/web/csdl/doi/10.1109/MC.2008.468

[15] The UK Research Data Service Feasibility Study, UKRDS, Report and Recommendations to HEFCE, 19th December 2008, http://www.ukrds.ac.uk/

[16] Comparative analyses on clinical research infrastructures, networks, and their environment in Europe, http://www.ecrin.org/index.php?option=com_docman&task=doc_download&gid=11&Itemid=68

[17] http://www.jisc.ac.uk/

[18] http://www.rin.ac.uk/

[19] http://www.casparpreserves.eu/

[20] http://www.clarin.eu/

[21] http://www.dans.knaw.nl/en/over_dans/

[22] http://www.textgrid.de/en/startseite.html/

[23] http://www.mpi.nl/DOBES/

[24] http://www.hrelp.org/

[25] http://www.dariah.eu/

[26] http://www.cessda.org/

[27] http://www.share-project.org/

[28] http://www.compare-project.org/

[29] http://www.esds.ac.uk/

[30] http://www.dcc.ac.uk/

[31] http://www.dpconline.org/graphics/index.html

[32] http://www.icpsr.umich.edu/ICPSR/

[33] http://www.ncess.ac.uk/

[34] http://www.elixir.europe.org/page.php?page=home

[35] http://www.eatris.eu/

[36] http://www.ecrin.org/

[37] http://www.fmp-berlin.de/eu-openscreen.html

[38] http://www.infrafrontier.eu/

[39] http://www.bbmri.eu/

[40] http://www.ebi.ac.uk/

[41] http://www.biosapiens.info/

[42] http://www.embracegrid.info/page.php?page=home

[43] http://www.emmanet.org/index.php

[44] http://www.eumodic.org/

[45] http://www.bodc.ac.uk/

[46] http://www.opendoar.org/

[47] http://www.driver-community.eu/ and http://www.driver-support.eu/

[48] http://www.metaforclimate.eu/

[49] http://www.d4science.eu/

[50] http://wiki.services.eoportal.org/tiki-index.php?page=HMA%20Wiki

[51] http://www.gmes-geoland.info/index.php/

[52] http://www.myocean.eu.org/index.php/project/

[53] http://www.insea.info/

[54] http://www.mersea.eu.org/

[55] http://www.parse-insight.eu/

[56] http://wiki.services.eoportal.org/tiki-slideshow.php?page=SOSI%20Wiki&slide=1

[57] http://grelc.unile.it:8080/ClimateG-DDC/

[58] http://www.eu-degree.eu/

[59] http://www.seadatanet.org/

[60] http://ec.europa.eu/maritimeaffairs/emodnet_en.html

[61] http://www.envirogrids.net/

[62] http://www.insa-vlabs.org/hiddra

[63] http://www.genesi-dr.eu/

[64] http://dublincore.org/

[65] http://www.lifewatch.eu/

[66] http://www.biocase.org/

[67] http://www.gbif.org/

[68] http://www.e-taxonomy.eu/

[69] http://www.enbi.info/forums/enbi/index.php

[70] http://www.eur-oceans.info/EN/home/index.php

[71] http://www.marbef.org/

[72] http://www.marine-genomics-europe.org/

[73] http://www.synthesys.info/

[74] http://www.eu-orchestra.org/

[75] David J McKillen et. al., Marine Genomics: A clearing-house for genomic and transcriptomic data of marine organisms, http://www.biomedcentral.com/1471-2164/6/34

[76] http://www.alliancepermanentaccess.eu/

[77] http://www.delos.info/

[78] http://www.insea.info/documents/insea/downloads/d4_2_metadatahandbook.pdf

[79] http://www.gmes.info/

[80] http://www.health-e-child.org/

[81] http://www.iassistdata.org/

[82] http://www.rcsb.org/

[83] http://www.euro-vo.org/

[84] http://www.ariadne.ac.uk/issue58/dplta-hep-rpt/

[85] http://ec.europa.eu/information_society/activities/health/policy/index_en.htm

[86] http://www.euro-vo.org/

[87] Mandl KD, Szolovits P, Kohane IS (February 2001). "Public standards and patients' control: how to keep electronic medical records accessible but private". BMJ 322 (7281): 283–7. http://www.bmj.com/cgi/content/full/322/7281/283?view=long&pmid=11157533

[88] Ruotsalainen P, Manning B (2007). "A notary archive model for secure preservation and distribution of electrically signed patient documents". Int J Med Inform 76 (5-6): 449–53. http://www.ncbi.nlm.nih.gov/pubmed/17118701

[89] http://icos-infrastructure.ipsl.jussieu.fr/index.php?p=hom

Chapter 2

Metadata and quality of data

1 INTRODUCTION

1.1 PURPOSE

This chapter aims to describe issues pertaining to metadata and quality of data resources for consideration by the e-IRG. The intended audience for this document include the e-IRG delegates and the e-IRG Data Management Task Force (DMTF).

1.2 SCOPE

This chapter is restricted to issues relevant to metadata and quality of data resources.

1.3 ORGANIZATION

The chapter is organized as follows: it starts with a general introduction, then considers issues relating to metadata, and finally examines issues relating to the quality of data resources.

1.4 OVERVIEW

This chapter was created as part of the activities of the e-IRG Data Management Task Force [1] and is meant to describe basic principles and requirements for the metadata descriptions [1] and the quality of resources that will be stored in accessible repositories as described by the ESFRI Working Group about Digital Repositories [2]. The principles and requirements specified in this document should be considered by all research infrastructures as baselines, since they are discipline independent.

2 METADATA

2.1 INTRODUCTION AND OVERVIEW

Recently Tony Hey used the term Data Intensive Science as the 4th Research Paradigm [3] to describe an essential paradigm shift in many research areas. Almost all disciplines are confronted with the Digital Data Deluge and therefore need to find ways to manage its records and to solve the retrieval problem. Descriptive metadata (DM) describe resources and services with the help of useful key-value pairs and in so far classify them according to a number of specified semantic dimensions. DM also adds information that is not part of the resource or service

[1] Metadata in the broad sense is data about data, including any form of annotations on resources or resource fragments. In this report metadata is used in the restrictive definition of descriptive keyword based metadata that is meant for discovery purpose for example. It can be seen as the electronic versions of the "old" library cards.

such as documenting its creation and modification contexts, its internal encoding and formatting principles, its availability and accessibility etc. At the recent conference of the Alliance for Permanent Access [4] the need for high quality metadata for long-term preservation objectives was confirmed, i.e. DM should also include provenance information to allow people to trace back with help of which automatic operations the resource was created. Since DM represents resources or collections of resources, contains important information about their nature it plays an increasingly important role in a eScience scenario.

DM is not new of course. Libraries dealing with large numbers of books are very familiar with this concept in form of library cards which can be used by the librarians and the users to inform themselves and to search for the books. Not surprisingly important initiatives came from the (digital) library world to introduce electronic cards for all types of web resources. The Dublin Core element set [5] is a result of the corresponding discussions. Slightly delayed the disciplines who were confronted with large amounts of different data and data types and who wanted to share data at a large scale started thinking about element sets that could be used for their management and retrieval. Naturally most of the disciplines decided to use their terminology when the were defining their classification system and due to their more precise domain knowledge the semantics of their elements where defined much more precise than those of the 15 core elements of Dublin Core.

This document does not want to be comprehensive but just refer to a few sets that have been defined by certain communities. The community dealing with learning objects created the LOM set [6] which is an elaborated structured set. In the linguistic community two sets were defined: (1) The OLAC set [7] oriented itself at the Dublin Core set, refined their semantics and added a few elements. (2) The IMDI set [8] started from domain terminology and the requirements of the various linguistic data types. The climate researchers re-used and extended the ISO 19115 set to create a Common Information Model (CIM) to describe climate models, data and numerical experiments[9]. The astronomers have defined their own set [10] to be able to generate the Virtual Space Observatory. In the humanities the TEI header elements [11] are widely used. Finally in the cultural heritage domain the CIDOC CRM conceptual reference system [12] covering more than just metadata elements is being used. Many more examples could be given.

Where metadata is used in dynamically changing research environments it became obvious during the last decade that flexibility is one of the most important requirements: (1) New elements are defined and need to be added; (2) researchers are using different selections of the defined elements to prevent overhead and (3) researchers want to borrow elements from different sets to re-use already existing definitions. Therefore, in addition to the fixed schema solutions we can already indicate a few communities that work on flexible component based schemas [13] where the elements definitions are kept separately in openly accessible category registries [14]. This view is widely shared by the Dublin Core community that is restricting itself more or less to maintain the definitions of its core set and its set of semantically more specific qualified elements.

In the following we will address a number of essential topics with respect to metadata such as usage, scope, persistence, aggregations, standardization, interoperability and quality.

2.2 USAGE

Descriptive metadata (DM) is used by researchers primarily to find proper resources and services and by repositories for resource management purposes. However, in the eScience scenario new types of usages will emerge since DM can be seen as “representations” of resources and services, i.e. they will be subject of research questions and they will be used by automatic agents for various kinds of operations such as profile matching. To fulfil these functions they need to incorporate keyword-value pairs that can be used for querying, quick inspection, identification, for grouping resources and for referencing them.

Providing metadata describing any kind of research resources and services is an urgent requirement for service providers and resource repositories.

2.3 SCOPE

DM therefore in general cover keywords that are both domain specific and domain unspecific. Yet we are not that far that a vocabulary has been agreed across all domains for the unspecific parts. Thus the currently used element sets are using domain specific terminology to a large extent to ensure its usefulness within the domain. Discipline

crossing initiatives such as Dublin Core for example are intended to allow users to describe web resources with the help of one single element set the semantics of which obviously are defined very broadly. Since in the research domain metadata descriptions are part of research queries, only specific domain-oriented terminology will be sufficient. However, there will also be users that are not interested in specific results, but in discipline-crossing overviews.

There is an increasing pressure for disciplines to agree on a set of semantically specific enough elements that allows researchers to describe their services and resources.

2.4 PROVENANCE

Increasingly often resources are the result of manipulating other resources with the help of automatic tools such as transformers etc. For proper interpretation, long-term preservation (prevent concatenation effects) and further processing it is of greatest importance to maintain provenance information about the creation history. DM should either contain this information itself or refer to a provenance object.

DM should include or refer to provenance information to support long-term preservation and further processing.

2.5 PERSISTENCE

DM refer to the resources and services they describe. In the emerging eScience scenario these references need to be made of persistent identifiers. The DM themselves need to be identified by persistent identifiers due to their extremely important role and repositories need to take care of their long-term persistence.

Metadata descriptions need to be persistent, to be identified by persistent identifiers and also to refer to the resources and services they represent by using persistent identifiers.

2.6 AGGREGATIONS

DM can describe data at different levels of aggregations: They can be used (1) to represent bundles of resources that have a strong internal relation; (2) to represent any types of collections defined by the intentions of the creators and (3) to represent virtual collections that are defined across repositories or service providers. In many of these cases metadata descriptions will make these aggregations citable.

Descriptive metadata have an enormous potential to describe various forms of groupings and can give them an identity, i.e. making them citable.

2.7 STANDARDIZATION

For DM it is of great importance that they are standardized to allow anyone to operate on them. This means that the elements and their values need to be defined. It also means that the format needs to be schema based. Due to the various intentions and contexts DM need to be very much tailored to the specific needs, i.e. researchers want to use selections of element sets and even want to add new elements. To cope with these requirements, we can see a transition towards using flexible components that make use of registered concepts [13]. Concept registries based on ISO 12620 such as they are specified by ISO TC37/SC4 [14] may form practical solutions.

Descriptive metadata needs to be based on well-defined element semantics and a schema-based format to cater for presentations for humans and machine operations. Where fixed schema solutions are given up, elements need to be re-used which are registered in open registries.

2.8 INTEROPERABILITY

DM needs to be open to all users and service providers to allow all researchers resource and service discovery even in an interdisciplinary framework. Yet there are no universally agreed methods to register the categories used and

for mapping them [2]. DM offered by data providers will be harvested by service providers. Currently, the OAI Protocol for Metadata Harvesting [15] is used for harvesting, which also requires the provision of Dublin Core records, i.e. a mapping from the discipline semantics to the DCMI semantics needs to be provided. In addition the DM records can be offered and the service providers can interpret them, since schema and element semantics have been clearly defined. In some domains DM is offered in a protocol neutral fashion by offering the XML objects via the web. The available XML schema allows service providers harvesting these DM objects to interpret them.

Descriptive metadata needs to be open and offered for harvesting via widely accepted mechanisms to cater for interdisciplinary usage.

2.9 Quality

Due to the extremely increasing amounts of data there needs to be an increasing pressure on researchers to produce high quality metadata descriptions and therefore to document the resources and services it describes. Providing proper citation, resource history and authenticity information will become increasingly relevant.

Researchers need to be urged to produce high quality metadata descriptions.

2.10 Earliness

DM cannot be created without costs. It is known that if metadata is not created immediately at resource creation time the costs will increase rapidly [16] and the quality decreases requiring costly curation efforts. Therefore resource creation tools should support the immediate creation of metadata, but not all supplementary information can be created automatically.

Researchers should be motivated to create metadata immediately and tool developers should add those descriptors that can be created automatically.

2.11 Availability

Due to the importance of DM in various ways as described above it should become a MUST for researchers to create DM and for infrastructure and tool developers to provide means to automatically at quality metadata. These metadata descriptions MUST be made available in a persistent way together with the resources and services they describe.

It is MUST for all resource and service providers to create and provide quality metadata descriptions.

3 Quality of Data Resources

3.1 Introduction

The quality of research is usually measured in terms of scientific output. The quality of scientific output, especially in paper journals, has traditionally been assured through the system of peer review. For electronic publications, new ways of reviewing articles in digital repositories have been sought and implemented since the mid 1990s [1]. Applications for projects are also usually peer reviewed. According to National Institutes of Health (NIH) in the US, the increasing breadth, complexity, and interdisciplinary nature of modern research has necessitated a more formal review of the NIH peer review system. A recent report identifies the most significant challenges and proposed recommendations that would enhance the peer review system [2].

The data bases underlying scientific publications are only rarely reviewed, although there is an increasing number of journals requiring the submission of such data sets in (publicly accessible) data repositories. Such journals have been labelled as "DAP-Journals' [3] or journals with a Data Availability Policy. Data archives, which have been set up since the 1960s, have always used the potential for checking of possible errors in data collections as a motive for their (continued) existence. Such digital data archives are the main advocates of quality assurance

[2] It is well-understood that it is impossible to create a universal ontology, since semantic mapping mainly has to do with the purpose of the task at hand. Some service providers such as in the cultural heritage domain create specific mappings focusing on a specific user group. Other such as the Max-Planck-Society will go the way to register all elements that are used and to allow users to easily manipulate mappings.

for research data. Quality control by data archives is usually achieved by painstaking and labour intensive checks on the data, carried out by data archive staff. Quality checks carried out include:

- check the format of the data files
- check whether a complete code book is available for coded data
- check the anonymity of personal data; data are de-identified by expunging names, addresses, etc.
- checks on missing values and overall completeness / data integrity
- consistency checks

Moreover, the description of the data sets by adding metadata (for archiving and retrieval) is often carried out by data archive staff, with some exceptions. Few archives use a self-deposit system, in which the depositors (i.e. the researchers who have produced the data) add the metadata describing the data they submit. In these cases, the data archives still perform (marginal) quality checks on the metadata and data deposited.

Some data archives have plans for introducing a form of review by users (which are normally peers in research data archives) similar to product or service reviews that are common in many web shops. Such review systems have to our knowledge not yet been implemented.

Internationally, there are several initiatives for setting criteria to certify digital repositories, such as TRAC [4], DRAMBORA [5] and the *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive* developed by NESTOR [6]. Both concentrate on quality of digital archives as organisations and (technical) service providers, not necessarily on the quality of their contents. The content of research data repositories is always dependent what is submitted by depositors, researchers, who are primarily responsible for the quality of their work.

The OECD has published a set of thirteen principles and guidelines for access to research data from public funding [7], among which several are linked to data quality, and one is explicitly labelled "quality". The guidelines concern: *A. Openness, B. Flexibility, C. Transparency, D. Legal conformity, E. Protection of intellectual property, F. Formal responsibility, G. Professionalism, H. Interoperability, I. Quality, J. Security, K. Efficiency, L. Accountability, M. Sustainability*. In the guidelines improved access itself is seen as benefiting the advancement of research, boosting its quality (p. 8). The guideline on quality per se is reproduced in Table 2.1.

Note: italics and bold added by the author

In a recent report by the Strategic Committee on Information and Data (SCID) of the International Council for Science (ICSU), an advice is formulated on the future organization and direction of the activities in relation to scientific data and information. Although the term "data quality" appears 17 times in the text, it is not explicitly stated what quality means or how it can be guaranteed. The report makes it very clear, however, that data organisations are to play an important role in ensuring the quality of and access to research data, as in Table 2.2.

Note: bold and italics added by the author

3.2 SHARING DATA AND QUALITY ASSURANCE

In a recent report commissioned by the Research Information Network (RIN), one chapter is devoted to "Quality assurance in the data creation process" [8]. With regard to creating, publishing and sharing datasets the RIN report identifies three key purposes:

1. datasets must meet the purpose of fulfilling the goals of' the data creators' original work;
2. datasets must provide an appropriate record of the work that has been undertaken, so that it can be checked and validated by other researchers;
3. datasets should be discoverable, accessible and re-usable by others.

The value and utility of research data depends, to a large extent, on the quality of the data itself. Data managers, and data collection organisations, should pay particular attention to ensuring compliance with explicit quality standards. Where such standards do not yet exist, institutions and research associations should engage with their research community on their development. Although all areas of research can benefit from improved data quality, some require much more stringent standards than others. For this reason alone, **universal data quality standards are not practical.** Standards should be developed in consultation with researchers to ensure that the **level of quality and precision meets the needs of the various disciplines**. More specifically:
• Data access arrangements should describe good practices for methods, techniques and instruments employed in the collection, dissemination and accessible archiving of data to enable **quality control by peer review and other means of safeguarding quality** and authenticity.
• The **origin of sources should be documented** and specified in a verifiable way. Such documentation should be readily available to all who intend to use the data and incorporated into the metadata accompanying the data sets. Developing such metadata is important for enabling scientists to understand the exact implications of the data sets.
• Whenever possible, access to data sets should be **linked with access to the original research materials**, and copied data sets should be linked with originals, as this facilitates validation of the data and identification of errors within data sets.
• Research institutions and professional associations should develop appropriate practices with respect to the **citations of data** and the recording of citations in indexes, as these are important indicators of data quality.

Table 2.1: The quality guideline from the OECD principles and guidelines (2007)

Following an earlier priority area assessment exercise in this area, ICSU's declared strategic goal is: to facilitate a new, coordinated global approach to scientific data and information that ensures **equitable access to quality data and information** for research, education and informed decision-making [30].
The ICSU plans to create a new World Data System (as an ICSU Interdisciplinary Body), incorporating the existing World Data Centers (WDC) and the Federation of Astronomical and Geophysical Data analysis Services (FAGS) as well as other state-of-the-art data centres and services. This new structure or system must be designed to ensure the long-term stewardship and provision of **quality-assessed data** and data services to the international science community and other stakeholders.
ICSU has an important responsibility on behalf of the global scientific community for promoting the optimal stewardship and policy development for scientific data and information. Three major trends in data and information management are dramatically changing science. The first is the major step change in the sheer volume and diversity of data suitable for science. Many fields from geo-demographics to particle physics are witnessing dramatic increases in data and information volumes. The second is the availability of new information and communication technologies, such as Grid computing or Sensor Web, which means that very ambitious modelling and data processing are within the scope of an increasing number of scientists. The third is the increasing need for **scientific datasets to be properly identified, quality-assured, tracked and accredited** (for example, through assignment of digital object identifiers or DOIs). This requires professional data management and, in some areas, may involve review and publication of datasets. Publication and accreditation can also act as an important incentive for primary data producers to make their data available.
There is a need for global federations of professional state-of-the-art data management institutions, working together and exchanging practices. Such federations can **provide quality assurance** and promote data publishing, providing the backbone for the development of a global virtual library for scientific data.

Table 2.2: Excerpts from the Final Report to the ICSU Committee on Scientific Planning and Review [30]

Fulfilling the first and second of these purposes implies a focus on scholarly method and content; the third implies an additional focus on the technical aspects of how data are created and curated. The scientific or scholarly value of datasets that are not accessible for re-use by others can obviously not be assessed by independent peers.

The RIN report distinguishes data sets created by machines (such as telescopes, spectrometers, gene sequencers) from those created in other ways (such as social surveys, databases created by manual input, source editions of texts, etc.). This distinction roughly (though by no means not completely) coincides with the distinction between the sciences and the humanities. Machines that create data often have inbuilt data validation mechanisms. Manual checking is usually added, and in those disciplines where data are collected by other means manual verification may involve very detailed work.

Although there is no information on how many data sets are checked by others than the researcher herself, it is in many cases taken for granted that when a paper is accepted for publication after peer review, the underlying data will pass the quality standard as well. Peer review may involve checks of the supporting data. In some disciplines, reviewers do checks on data. In other cases, checking is superficial or absent, because the data are too complex or voluminous to be judged satisfactorily. Most researchers take other researchers' outputs on trust in terms of data quality and integrity. Moreover, there are no apparent signs of dissatisfaction with this state of affairs.

Summary:	17. Most researchers believe that data creators are best-placed to judge the quality of their own datasets, and they generally take other researchers' outputs on trust in terms of data quality and integrity. 18. There is no consistent approach to the peer review of either the content of datasets, or the technical aspects that facilitate usability. 19. Data centres apply rigorous procedures to ensure that the datasets they hold meet quality standards in relation to the structure and format of the data themselves, and of the associated metadata. But many researchers lack the skills lack the skills to meet those standards without substantial help from specialists.
Recommendation:	9. Funders should work with interested researchers, data centres and other stakeholders to consider further what approaches to the formal assessment of datasets – in terms of their scholarly and technical qualities – are most appropriate, acceptable to researchers, and effective across the disciplinary spectrum.

Table 2.3: Summary and recommendation with respect to data quality in the RIN report

The accessibility of data for re-use renders checking possible at a later time. Experimental, machine-produced data can in principle be re-created if the whole experiment is done again. In practice, such validity-checking procedures are rarely carried out. Nevertheless, new experiments with better measuring equipment, or the secondary analysis of survey data or text corpora, re-appraisal of digital scholarly editions and so on, may result in the discovery of earlier flaws, and in extreme cases in the exposure (and shaming of) mistakes or even fraud.

Although it is useful to distinguish between the scientific/scholarly content of data and the technical merits that facilitate re-use, it is questionable whether the two can be separately reviewed, as recommended in a report commissioned by the Arts and Humanities Research Council [9]. British researchers generally support the idea of instituting a formal process for assessing the quality of datasets, although they have concerns whether it will work effectively in practice. Among these are the difficulty to find reviewers who are willing and who have the expertise to understand and appraise the data [10]. Another concern involves the costs (in terms of money and time) of a formalized data review process.

It is not likely that the pressure to improve the quality assurance process for datasets will come from the researchers, although they generally seem to favour a more thorough assessment. Research funders, who are investing heavily in the data infrastructure, are in a better position to take the initiative to introduce a formal assessment process. This would also imply that data creation itself shall be rewarded with scholarly merit and scientific credits [11]. Funding agencies do not always require a formal data plan, although for certain subsidies in some countries, such a requirement exists. For instance, researchers receiving a grant from the Dutch funding

organisation NWO for a data creation project are required to sign a "data contract" with DANS [12], in which they are obliged to comply with the "Data Seal of Approval".

Whilst datasets that are deposited at data centres must conform to certain quality standards, there is no such imperative yet for researchers who look after their own datasets. According to the RIN study, some researchers believe this to be outside the boundaries of their research function, while others lack the skills (and/or time) to publish their data such that it can be discovered, accessed and re-used by the scholarly community.

3.3 ASSESSING THE QUALITY OF RESEARCH DATA

Recently a method for assessing the quality of research data, called the "Data Seal of Approval" (DSA) has been developed by DANS in The Netherlands. The ambition of the DSA is to ensure that research data of a guaranteed quality can be found, recognized and used in a reliable way [13]. An international board consisting of data archivists and researchers from various disciplines and countries has taken on the responsibility for the guarding of the DSA and its application in practice. An assessment procedure has been developed and tested and is currently undergoing its wider implementation and roll out. The quality guidelines formulated in the DSA are of interest to researchers and institutions that create digital research files, to organizations that manage, archive, curate and disseminate research files, and to users of research data.

This DSA contains a total of 16 guidelines for the application and verification of quality aspects with regard to creation, storage and (re)use of digital research data (see Appendix A). Although originally developed for application to the social sciences and humanities, the draft text of the DSA has also met with positive responses from the natural sciences. Slight modifications have been made to make the DSA more generically applicable. The international board will take additional requests for adaptations into consideration if these are required for certain research fields. However, the DSA guidelines seem well capable of serving the quality requirements of many disciplines, and even of areas outside of science (e.g. cultural heritage, public archives).

The criteria for assigning the seal of approval to data are in accordance with, and fit in with, national and international guidelines for digital data archiving such as *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive* [15] as developed by NESTOR, *Digital Repository Audit Method Based on Risk Assessment* (DRAMBORA) [14] published by the Digital Curation Centre (DCC) and Digital Preservation Europe (DPE), and *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist* of the Research Library Group (RLG). The *Foundations of Modern Language Resource Archives* of the Max Planck institution [14] and the *Principles and guidelines for the stewardship of digital research data* published by the Research Information Network [15] have also been taken into consideration. The DSA guidelines can be seen as a minimum set distilled from the above proposals.

Digital research data must meet five quality criteria:

1. The research data can be found on the Internet.
2. The research data are accessible, while taking into account ruling legislation with regard to personal information and intellectual property of the data.
3. The research data are available in a usable data format.
4. The research data are reliable.
5. The research data can be referred to.

The DSA guidelines focus on three operators: the *data producer*, the *data repository* and the *data consumer*.

- The *data producer* is responsible for the quality of the digital research data.
- The *data repository* is responsible for the quality of storage and availability of the data: data management.
- The *data consumer* is responsible for the quality of use of the digital research data.

Data producers

The quality of digital research data is determined by:

- Their intrinsic scientific quality;
- The format in which the research data and supporting information are stored;
- The documentation (metadata or contextual information) regarding the research data.

Scientific quality criteria indicate to what degree the research data are of interest to the business of science. The assessment by experts, colleagues in the field, is the main decisive factor for the scientific quality of research data. Three questions must be answered to be able to provide an assessment.

1. Are the research data based on original work performed by the data producer (researcher or institution that makes the research available) and does the data producer have a solid reputation? This question can be answered by providing information regarding the researcher and/or research group and by providing references to publications pertaining to these particular research data.
2. Was data creation carried out in accordance with prevailing criteria in the research discipline? The answering of this question requires information on the used methods and research techniques, including those for data collection, digitization or other means of data creation.
3. Are the research data useful for certain types of research and suitable for reuse? The answer requires information regarding the data *format, content and structure*. The data producer therefore provides sufficient information to enable fellow scientists to assess the research data.

Data format The bits that form a digital research object are organized according to the rules for a particular data format. Various data formats exist for digital objects. For all formats, there is a risk that they may become obsolete. That creates a chance that the data object may become unusable. For storage of data objects, so-called preferred formats are therefore used. Preferred formats are formats designated by a data repository for which it guarantees that they can be converted into data formats that will remain readable and usable. Usually, the preferred formats are *de facto* standards employed by particular research communities.

Documentation The data producer provides the research data with contextual information (metadata). There is a distinction into descriptive, structural and administrative metadata. These must be provided in accordance with the applicable guidelines of the data repository.

- Descriptive metadata are data required to be able to find research data and that add transparency to their meaning (definition and value) and importance. Examples of descriptive metadata are the data elements of the Dublin Core Element Set [16], with fields such as creator, type, and date.
- Structural metadata indicate how different components of a set of associated data relate to one another. These metadata are needed to be able to process the research data. When data are coded, the codebook will be a component of the structural metadata.
- Administrative metadata are required to enable permanent access to the research data. This concerns the description of intellectual property, conditions for use and access, and so-called preservation metadata needed for durable archiving of the research data.

Data repositories

The data repository is responsible for access and preservation of digital research data on the long term. Two factors determine the quality of the data repository:

- The quality of the organizational framework in which the data repository is incorporated (organization and processes);
- The quality of the technical infrastructure of the data repository.

Organization and processes Organizations that play a role in digital archiving and are establishing a Trusted Digital Repository (TDR) minimally possess a sound financial, organizational and legal basis on the long term. Depending on the task assigned to an organization, a TDR may distinguish itself qualitatively by carrying out research and by cooperating with other organizations in the realm of data archiving and data infrastructure. The outcomes of such research are shared, both nationally and internationally. In addition, these organizations will also share physical infrastructures, software and other knowledge among each other, where possible.

Technical Infrastructure The technical infrastructure constitutes the foundation of a Trusted Digital Repository. The OAIS reference model, an ISO standard, is the de facto standard for using digital archiving terminology and defining the functions that a data repository fulfils [17].

Data consumers

The quality of the use of research data is determined by the degree to which the data can be used without limitation for scientific research by the various target groups, while complying with certain rules of conduct. The open and free use of research data takes place within the legal frameworks and the policy guidelines as determined by the relevant (national) authorities.

The 'OECD Principles and Guidelines for Access to Research Data from Public Funding' [18] provide policy guidelines regarding access to research data, which are accepted by the governments of the OECD countries. The principles of 'Open Access' are moreover described in the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities [19], which are signed by over 250 scientific organisations in more than 30 countries (end of 2008).

In the Netherlands, the *Code of Conduct for Research* [20] is of importance for the use of research data. This Code of Conduct utilizes five core concepts that are essential for the quality of scientific research: due care, reliability, verifiability, objectivity, and independence. An additional *Code of Conduct for the use of personal information in scientific research* [21] focuses on responsible handling of privacy-sensitive data and describes the legal frameworks for scientific work with such data. These codes of conduct comply with the Dutch Personal Data Protection Act (WBP). This law provides the frameworks within which personal information may be used in The Netherlands. The Dutch Data Protection Board (CBP) monitors compliance with legislation that regulates the use of personal information.

4 METADATA AND QUALITY TASK FORCE MEMBERS:

Name	Organization	Initiative
Patrick Aerts	NWO	HPC
Hilary Beedham	Univ. Essex	CESSDA
Tobias Blanke	Kings College	DARIAH
Victor Castelo	Ministry of Science and Innovation, Spain	CSIC
Peter Doorn	DANS	DARIAH
Luigi Fusco	ESA	ESA
David Giaretta	STFC	PARSE.Insight
Maria Koutrokoi	Ministry of Development, Greece	NCP-RI
Diego Lopez	RedIRIS	EGEE,GEANT
Pasquale Pagano	ISTI	EGEE
Dany Vandromme	RENATER	ESFRI
Peter Wittenburg	MPG	CLARIN
Andrew Woolf	STFC	STFC
Matti Heikkurinen	Emergence-Tech	DMTF Support
Michele Landes	RENATER	DMTF Support

Table 2.4: DMTF-QUALITY Membership

5 REFERENCES

5.1 METADATA:

[1] http://www.e-irg.eu/

[2] ftp://ftp.cordis.europa.eu/pub/esfri/docs/digital_repositories_working_group.pdf/

[3] http://www.alliancepermanentaccess.eu/index.php?id=3

[4] http://colab.mpdl.mpg.de/mediawiki/EScience_Seminar_2009/EScience-Seminar_Fashion_Wave_or_Driving_Force_of_Progress#Contributions

[5] http://dublincore.org/

[6] http://de.wikipedia.org/wiki/Learning_Objects_Metadata

[7] http://www.language-archives.org/

[8] http://www.mpi.nl/IMDI/

[9] http://metaforclimate.eu/

[10] http://www.astrogrid.org/

[11] http://www.tei-c.org/index.xml

[12] http://www.datasealofapproval.org/

[13] http://www.clarin.eu/specification-documents

[14] http://www.isocat.org/

[15] http://www.openarchives.org/OAI/openarchivesprotocol.html

[16] http://www.alliancepermanentaccess.eu/index.php?id=3

5.2 QUALITY

[1] Harnad, S. (1996) Implementing Peer Review on the Net: Scientific Quality Control in Scholarly Electronic Journals. In: Peek, R. & Newby, G. (Eds.) Scholarly Publication: The Electronic Frontier. Cambridge MA: MIT Press. Pp. 103-108.

[2] http://enhancing-peer-review.nih.gov/

[3] Tine de Moor & Jan Luiten van Zanden (2008), Do ut des (I give so that you give back): collaboratories as a new method for scholarly communication and cooperation for global history. Historical Methods, Volume 41, Number 2 / Spring 2008, pp. 67 – 80. DOI: 10.3200/HMTS.41.2.67-80. See http://www.iisg.nl/publications/do-ut-des.pdf for a pre-print.

[4] Trustworthy Repositories Audit and Certification, see: http://www.oclc.org/research/announcements/2007-03-12.htm

[5] Digital Repository Audit Method Based on Risk Assessment, see: http://www.repositoryaudit.eu/

[6] See: http://edoc.hu-berlin.de/docviews/abstract.php?id=27249

[7] See: http://www.oecd.org/dataoecd/9/61/38500813.pdf

[8] Michael Jubb (2008) To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. Report commissioned by the Research Information Network (RIN). See: http://www.rin.ac.uk/data-publication

[9] Peer review and evaluation of digital resources for the arts and humanities, Arts and Humanities Research Council, September 2006. See: http://www.britac.ac.uk/reports/peer-review/index.html

[10] In a recent study of researchers' attitudes to peer review, 40% of reviewers and 45% of journal editors said it was unrealistic to expect peer reviewers to review authors' data. Ware, M (2008), Peer Review: benefits, perceptions and alternatives. See: http://www.publishingresearch.net/documents/PRCPeerReviewSummaryReport-final-e-version.pdf

[11] Although off-topic here, there is no reason why other digital products, such as scientific software, could not be treated in a similar way.

[12] DANS - Data Archiving and Networked Services – is an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NWO). DANS has the mission to keep research data in the arts and humanities and social sciences permanently accessible. See: http://www.dans.knaw.nl/en/

[13] See: http://www.datasealofapproval.org/

[14] Peter Wittenburg, Daan Broeder, Wolfgang Klein, Stephen Levinson, of the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands, and Laurent Romary of the Max Planck Digital Library in Munich, Germany. See: http://www.lat-mpi.eu/papers/papers-2006/general-archive-paper-v4.pdf

[15] See: http://www.rin.ac.uk/data-principles

[16] See: http://www.dublincore.org

[17] For further information on the OAIS reference model see: http://en.wikipedia.org/wiki/Open_Archival_Information_System

[18] OECD doc

[19] See: http://oa.mpg.de/openaccess-berlin/berlindeclaration.html

[20] See: http://english.vsnu.nl/web/show/id=88938/langid=42

[21] See: http://www.vsnu.nl/web/show/id=69988/langid=43/ (in Dutch; link not working on 29 June 2009)

[22] ICSU (2008) Ad hoc Strategic Committee on Information and Data. Final Report to the ICSU Committee on Scientific Planning and Review.

Chapter 3

Interoperability Issues in Data Management

1 INTRODUCTION

1.1 PURPOSE

This document aims at reviewing the interoperability issues in data management for consideration by the e-IRG [35]. The intended audience for this document include the e-IRG delegates and the e-IRG Data Management Task Force (DMTF).

1.2 SCOPE

The scope is restricted to interoperability issues in data management that are crucial to wider access to data content.

1.3 ORGANIZATION

The document is organized as follows: it starts with a general introduction, then surveys resource-level and semantic interoperability, looks at some examples of use cases, and examines miscellaneous related issues.

1.4 OVERVIEW

The key issue is to make scientific data reachable and useful for other scientific fields. As an example, take the SHARE project [181] (one of the Social Sciences and Humanities ESFRI [48]) dedicated to the ageing of the population. There are many aspects for that which are relevant to health, economy, politics and other fields addressing different scientific communities to justify cross-disciplinary access. It is likely that similar needs for interoperability are crucial for environment or biodiversity, etc. Interoperability in this document does not necessarily mean all data must be reachable and usable by all, as this may not scale in effort and cost, but **data should be widely accessible by design**.

The Need for Interoperability

As is described below, interoperability is often isolated to individual communities, driven directly by community-specific projects, standards, or to meet urgent needs. In the wider context, interoperability promises further advantages [108]:

- By promoting standard interfaces or protocols, interoperability can enable data sharing between communities;

- It can prevent users from being "locked in" to a single service provider;
- Conversely, a service provider providing interoperable interfaces can be more attractive to users, because users are more likely to pick a provider who will not "lock them in;"
- It can enable service providers to serve multiple communities with a single service implementation;
- It can enable customers to try out "more advanced" middleware stacks by giving them the assurance they can always revert to the "less advanced" stack because they interoperate;
- While it may seem self-evident, it is important that a service interoperates with itself (see Section 2.4).

Specifically, we highlight the need for open standards. Open standards – meaning that anyone is free to implement them, unencumbered by licences or patents, not just today but also in the future – are often essential to achieving interoperability in this wider context, beyond a single community. It is often important to have a standards body, not just a single company or institution, "owning" the standard: for example, much of the success of the Web is due to the HyperText Transfer Protocol (HTTP) being standardised in W3C. It is also often helpful if the standards group is open to participation: large membership fees can prevent smaller communities or academic institutions from participating in the standards process.

Alternative means of achieving interoperability include building adapter interfaces. Put simply, an adapter provides a different front end to a service.

Levels of Interoperability

The approach taken here is to differentiate between various levels of interoperability, where for the communities the two generic layers of (a) resource format and domain encoding, and (b) general semantic interoperability (how to describe this is uncertain), are the most important.

Resource-level Interoperability: As an example, CLARIN [18] is working on interoperability in linguistics. This domain includes language resources of all types: audio, video, time series (eye tracking, brain imaging etc), texts, annotations, lexica, ontologies, schemas, etc. It is collaborating with ISO [77] in this area to develop standards. The problem is as always: the need to be at a level of abstraction that covers all disciplines. From the many discussions of CLARIN with libraries and various communities in the recent months, it seems that interoperability solutions for resources structures and content are very much **discipline dependent** and **discipline driven**, i.e. there are as many standards as disciplines and even sub-disciplines. Of course there are cross-discipline standards such as MPEGx [125] codecs for video resources that are universally defined, but as soon as meaning is explicitly encoded the differences begin. Even XML [49] is just the structuring language for text documents; it is a schema that defines certain document classes and still it does not specify the semantics. A schema can refer to data category registries that include definitions of the tag sets used, etc. If two schema elements both refer to the same data category then, for example, semantic assumptions can be made. Still problems may be encountered if the semantic scope of a data category is loosely defined, such as in Dublin Core [34], and when it appears in different structural contexts - interpretation difficulties can occur.

General Semantic Interoperability: In CLARIN there were additional discussions about the question of how to establish an integrated and interoperable domain of language resources and (web) services. In collaboration with ISO TC37/SC4 [99] there is an effort to define so-called *pivot formats*, which actually are generic data models for specific structured linguistic data types such as annotated streams (texts, audio, video), lexica or data category registries. Generic models are more abstract specifications with a higher expressive power, i.e. they have some flexibility and can therefore represent a wider class of documents without giving up the essential characteristics of the linguistic data type. For many users flexibility, however, means additional complexity, so user interfaces need to hide this complexity. For data category registries (ISO-12620 [84]) and for lexica (LMF [117]) such generic models have been defined and in collaboration with ISO the first tools have been made available to support these generic models/formats (ISOcat [102], LEXUS [118]).

Syntactic versus Semantic Interoperability: At a later date it may be useful to further distinguish between syntactic and semantic interoperability. For example, in astronomy interoperability aspects are tackled by an international alliance (called International Virtual Observatory Alliance [78]). There are *Working Groups* dealing with the different aspects, and this syntactic/semantic distinction can be seen, for example, in the fact that there are separate *Query Language* and *Semantics* working groups.

2 Resource-level Interoperability

It is useful to distinguish resource-level interoperability in different layers: device, communications, middleware and deployment, as well as to distinguish between interoperation and interoperability.

2.1 Device Level

At the lowest level, the Storage Networking Industry Association (SNIA) [176] is the standards body for storage. It manages interfaces to the storage devices, and is working on standardisation of future developments.

2.2 Communications Level

Network Infrastructures Data intensive research involves large scale data transfers, which themselves require state-of-the-art network infrastructures that are capable of adapting flexibly to the needs of the applications and researchers relying on them.

Transfer Protocols: Data transfer protocols are standardised typically by the World Wide Web Consortium (W3C) [205] (HTTP [67]) or the Open Grid Forum (OGF) [145] (GridFTP [61][134]). There are numerous data transfer and access protocols for other types of storage middleware. For example, some researchers use the OAI-PMH protocol for metadata harvesting [141], the newly issued OAI-ORE [142] to serialise XML/RDF data, or container formats like METS [124] or MPEG21 [127] to package data for exchanges.

Web Services: Web services are important building blocks for many protocols in data management (and elsewhere). It is important to have libraries and implementations for different languages. Web services build on XML [49], SOAP [172], WSDL [207] and UDDI [199] (standardised in W3C), and higher level web services based protocols often standardised by OASIS [157], and interoperation between implementations is driven by the WS-Interoperability Organisation [206].

Data Movers: Within standards, discussion of interoperability issues in hardware data movers took a big step forward following the publication of Danny Cohen's classic 1981 paper *On Holy Wars and a Plea for Peace* [16] that categorized big-endian and little-endian architectures, e.g. Ethernet [45] enforces a big-endian network standard. But beware, see *Multiplexed Buses: The Endian Wars Continue* (1990) [107]. Many hardware bus standards subsequently devoted specific attention to data mover interoperability issues, e.g. Futurebus [54], SCI [171], USB [200] and Firewire [53]. Valuable lessons can be learnt by perusal of the related literature. There is a greater problem and less uniformity in data movers in software standards. For a small-scale example, consider MPI:

- *MPICH-1 [129] and MPICH-G2 [128]:* These libraries automatically convert data between incompatible architectures using a *reader-makes-right* model; that is, any necessary data conversion is done by the receiver. The libraries do not convert TCP messages to a *network standard* (i.e. a neutral format) before sending the message. The data conversion is done in the MPICH layer above any ADI (e.g. the Globus ADI), i.e. does not have to be reinvented for every new device.
- *MPICH2 [130]:* Does not provide data conversion (yet), hence it is not possible to run MPICH2 applications across heterogeneous architectures.

- *OpenMPI [151]:* OpenMPI does data conversion. The data conversion is done in the core layer rather than the MCA device, i.e. as for MPICH-1 and MPICH-G2, it does not have to be reinvented for every new device.

2.3 MIDDLEWARE LEVEL

There are many, many different storage systems, managing single disk to multi-petabyte tapestores. As a rule there are few standard interfaces, but many systems attempt to provide POSIX (IEEE 1003.1-1988 [161]) access. Applications can call the API to access and update data. The standards body for POSIX is IEEE [74]. The full POSIX standard itself is quite complex and therefore quite difficult to implement. It also enforces requirements on filesystems which rather than being absolute requirements would be better as ones which the filesystem could decide to meet or not. Relaxing some requirements could allow for improved performance in some cases (most NFS implementations do this).

Data storage One of the promises of storage middleware is to promote a single consistent interface to diverse underlying storage systems, thus providing interoperable storage. In Grid middleware there are two different Grid storage interfaces, the Storage Resource Manager (SRM [178]) which is a control protocol for accessing mass storage and an open standard in OGF, and SRB [177] which is a "data grid" implementation by SDSC. SRM has at least six different interoperating implementations, e.g., StoRM [179], DPM [30], and d-Cache [23].

In "cloud" storage, there are currently fewer known standards. A sensible approach may be to wait and see – eventually, a de facto standard may emerge.

File Catalogues and persistent identifiers: File Catalogues manage files within an information environment, and they are often synchronised with Metadata Catalogues (see next section). For example in data grid environments, certain files may be replicated to various locations in order to ensure the stability of the data.

Each data management software has their own file catalogue component, for example DIGS (formerly QCDgrid) [6], LFC [114][5], RLS [59]. There are no formal standards.

Interoperability across information environments and over long periods of time is achieved through Persistent Identifiers (PID). Some studies have shown that e.g. URLs are unstable, with a half-life of only 4.6 years [110]. Persistent identifiers allow the unique identification of data (depending on the identifier scheme, either only files or both database entries and files) across systems. Whenever a data item moves, the reference of the PID needs to be adapted, yet the PID as such does not change. There are various PID schemes [163], also community-specific ones. Amongst the most prevalent ones are the Uniform Resource Identifier (URI) [195], the Uniform Resource Name (URN) [196], and the Handles [65]. Layered service providers such as the Digital Object Identifier (DOI) [29] add functionality on top of schemes such as the Handles.

Metadata Catalogues: The basic principles and requirements for the metadata descriptions and the quality of resources that will be stored in accessible repositories are described in a companion e-IRG Data Management Task Force report on metadata and quality [123]. There are often no generic standards at this level. There are a large number of standards, sometimes from standards bodies, such as the Open Geospatial Consortium (OGC) [144] for geospatial disciplines, or more informally, such as the iCAT project [68] for neutron spallation sources and synchrotron sources and the Data Documentation Initiative (DDI) [24] for social sciences and humanities. But they are more often specific to communities. For example, in linguistics alone there are three major strands: (a) ISLE Meta Data Initiative (IMDI) [81] was designed by community experts from scratch using community specific terminology; (b) Open Language Archives Community (OLAC) [146] was defined as an extension to Dublin Core; (c) many researchers in the world of textual data are using Text Encoding Initiative (TEI) [184] header tags. There are some catalogue solutions and some gateways between the different sets. CLARIN [18] is working on a more generic model that can incorporate the various sets so that by referring to the data category registry semantic interoperability is achieved. In other disciplines, for example, climate researchers have agreed on some standards and a catalogue, astronomers have defined a standard for their *Registry of Resources* which includes the Dublin Core [34], with interfaces defined as a standard WSDL document, and harvesting also supported through the existing Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) [141], and in e-Learning those

working on Learning Object Metadata (LOM) [115] also have solutions. At the infrastructure level there are some relatively widely-used grid-enabled metadata services, e.g. AMGA [4].

Federated Databases: To achieve interoperability on a middleware level a solution is required to transparently integrate heterogeneous database systems. An integration standard has been proposed by the grid community, The Open Grid Services Architecture Data Movement Interface OGSA-DMI [25]. This will provide a standardized mechanism for moving data from its source to its destination, i.e. an abstract interface for data movement (point-to-point, third-party) that is transport-agnostic. In this way the complexity for moving data within a grid is greatly reduced.

Further interesting integration standards are being developed by the OGF Database Access and Integration Services Working Group (DAIS-WG) [27].

Digital Repositories: There is as yet no common name for this type of information component: repository, object store, knowledge management system, etc. The largest community working with object stores refers to them as "repositories", but one must distinguish between the system that provides the functionality and the repository itself. Repository systems manage digital objects, where an object is anything from a simple file to a complex object composed of various files, metadata, and relations with other objects and functionalities. Object stores are e.g. used in the medical field to manage all the resources associated with a patient, from X-Rays (image file) to diagnoses (text processing); for managing satellite data (image files); for publications (PDF, text processing); for archaeological data (images, videos, ultrasounds of the ground, etc); and digitisation centres.

Obviously, repository systems are complex software that are often built on semantic technologies like XML databases, RDF triple stores, etc. They combine file catalogues and metadata catalogues as mentioned in the previous sections. Well known repository systems include SRB [177] and iRODS [80], Fedora Commons [51], DSpace [33], ePrints [42], eSciDoc [43] and LAMUS [112].

For various communities and virtual research environment, repositories are the infrastructure of choice. Proper repositories are the basis of any reliable and persistent resource providing system. Existing standards with regard to "trusted digital repositories" aim to define technological and organisational criteria for repositories that have addressed the preservation of their digital objects in a trustworthy way (e.g. OCLC Certification [139], the Certification of Digital Archives [13], and the Data Seal of Approval [26]). Convergence between grid technologies, repositories, clouds, digital preservation and other infrastructure technologies is the objective of activities including the OGF Repository Working Group [135], DRIVER e-Publications [31] for linking research data with publications, and activities in the framework of the OpenRepositories community [152].

2.4 Deployment Level

Even with interoperable services, any new deployment always carries a risk of disturbing interoperability. With services that are not guaranteed to be interoperable (e.g. those involving proprietary interfaces, or incompatible changes in protocol versions) it is important to update the clients (i.e., tools, libraries, or applications accessing the service) at the same time as the server providing the service. Without automated tools, such synchronized updates are even more risky.

Interoperability thus plays an important role in reliable service provision: when we *do not* have interoperability within a particular service, not only do all clients have to be upgraded in step with the server, but all other servers in the infrastructure have to be upgraded at the same time – otherwise they will not be accessible by the clients, nor will they interoperate between other servers (e.g. for data transfers between the servers). In lieu of automated tools it is advantageous to upgrade service providers individually, out of step; in that case if there is a problem with the upgrade, it is experienced only at a single site rather throughout the whole infrastructure.

Likewise, interoperability between versions allows a service provider to downgrade, to roll back the upgrade to the previous version. This can be helpful for a service provider: although developers and service providers test upgrades before rolling them out, tests are of course not guaranteed to catch all problems. Again, any non-interoperability forces a synchronized downgrade.

There are few automated tools to handle these kinds of deployment scenarios, but for example, see the Transactional Deployment System (TDS) [192].

2.5 Interoperability versus Interoperation

The Open Grid Forum (OGF) [145] has during the last few years hosted activities known as Grid Interoperation Now (GIN) [133]. The aim was to see what could be achieved now, with little or no development effort. GIN was split into four areas, one of which was *gin-data* for data management. To some extent, this activity depended on *gin-auth* (authentication and authorisation) and *gin-info* (information systems). The fourth activity was on job submission. GIN distinguished between **interoperation** and **interoperability**: the latter is true interoperability as you would expect from e.g. storage devices that share protocols and interfaces. The former was the main goal of GIN, the short term fixes making things interoperate now, e.g. by bridging protocols or building adapters (but avoiding large development efforts). This is a very important distinction.

3 Semantic Interoperability

While resource-level interoperability is concerned with ensuring compatibility of implementations (hardware, software), semantic interoperability is rather concerned with enabling data and information flows to be understood at a conceptual level. We review below a number of approaches that have been taken to address semantic interoperability.

3.1 Interoperability through Data Integration: OM2

The OM2 [147] is an open source data management platform with several unique characteristics. Born as a tool for enterprise directory management (hence its name), it has evolved into a full featured data integration framework, supporting semantic routing and schema adaptation. The OM2 is based on a layered architecture with parts of the OM2 stack providing transport, message and application services. Central to the OM2 architecture is the use of formal data-model specifications in the form of semantic web ontologies represented in OWL-DL. OM2 can be thought of as a message-driven SOA for the semantic web.

An OM2 node is the basic building-block of the architecture. An OM2 node is a software component which sends or receives (or both) OM2 messages. An OM2 message is an RDF graph serialized as RDF/XML [170] which conforms to the OM2 message ontology. The OM2 is at the basic level an asynchronous system. Nodes in an OM2 deployment exchange messages with each other and process the contents of the messages, including routing decisions according to those contents. A message may be part of a chain of messages which provides local ordering of sequences OM2 messages. This should not be confused with the notion of a *connection* or *state* which does not exist.

The OM2 architecture is able to support a wide range of applications. The original motivation for the development of OM2 was to build a vendor- independent distributed identity management platform able to support near real-time updates. The basic OM2 architecture however does not include any dependencies on this particular application. The application layer can be thought of as the part of the OM2 node which does the actual message processing. This part depends on services provided by the message layer.

3.2 Interoperability through Ontology Support within Directories: COPA

COPA [19] is a coding schema that allows effective searches in hierarchical structures, at one level or recursively, and do not impose any relevant complication to existing storage and retrieval systems. The system is oriented to store in parallel different classes of metadata and it is specifically tailored to take advantage of LDAP-based systems.

The basic idea is the creation of string identifiers made of the concatenation of special sets of symbols, that incorporate hierarchy position data in them. These identifiers are added to the data available for a certain element (attributes in a directory object). The regular structure simplifies the definition of both positive and negative filters, making very easy the expression of complex searches inside the virtual hierarchy defined by the codes. The COPA coding scheme can be applied to any hierarchical naming structure, using the COPA code at the *leaf* part of the name, though the common practice is the use of *uniform resource names* (URNs).

One of the main features of COPA is that it allows the decoupling of the actual representation of the information from the view that is offered at a given time. A hierarchical classification in knowledge areas can be maintained on a flat structure or on a structure based on organizational criteria. The knowledge area hierarchy is kept by the metainformation using COPA codes. This way, the (possibly multiple) affiliation of an object to a certain area is represented in the object itself, not by means of its position in the hierarchy (as it is the common approach up to now). Keeping an actual flat structure in the information tree is key for maintaining the independence of entries with respect to the position of the objects related to them, so their position(s) within the structure may be changed without actually moving them in the supporting system. A change in the corresponding attribute(s) is enough.

Building ontologies, in which the relationships among areas are richer than the simple specialization/generalization provided by the hierarchy, can be accomplished by adding other attributes to the entries, each one modeling a different relationship. These attributes will hold the code of the areas related to the one represented by the entry. And this can, of course, be mixed with the multiple virtual views.

3.3 Interoperability through Simplicity: ARCA

ARCA [3] is the name of both a system for metadata exchange and a portal for data access developed by the RedIRIS [167] community. Originally conceived for harvesting and accessing metadata about multimedia content, it is quickly evolving into a general mechanism for sharing data about collections of arbitrary digital objects, and incorporating support for additional elements, like federated access management.

ARCA is based on the exchange of RSS [166] description for digital objects, that are grouped into *feeds* corresponding to specific collections, and conveyed through *channels* associated with data sources.

The use of a simple, flexible and extensible framework like RSS provides a simple yet powerful entry point for metadata exchange and aggregation, allowing for:

- The development of user-friendly, intuitive access interfaces
- The easy integration of heterogeneous data sources into a uniform browse and search system.
- A very low adoption threshold for repositories willing to participate.
- The parallel support of several ontologies.
- Different levels of detail in descriptions.
- A simple interconnection with other metadata formats and portals.
- The leverage of collaborative environments, both at the service (mash-ups) and user (Web 2.0) levels.

3.4 Interoperability through Transcoding and Metadata: VP-Core

VP-Core [202] is a state-of-the-art, scalable Middleware Media Distribution Platform that facilitates access to, and usage of (shared) storage capacity, metadata databases, transcoding- and streaming servers. The platform offers functionality for searching, playing, uploading, transcoding, as well as a fine granularity media access control system towards its users. VP-Core is based on the Representational State Transfer (REST) [169] architecture and is designed to support content streaming applications by providing a back-end-, audio- and video-infrastructure. In the context of DMTF this platform offers transcoding and metadata facilities that can be leveraged to achieve interoperability with other platforms.

To achieve codec interoperability, the REST-based service Application Programming Interface can be used to specify what transcoding is needed for interoperability with other formats. For example, video content can be uploaded in a high quality format such as MPEG2 and it can be retrieved in a low quality format such as flash video through an on-the-fly transcoding service. To achieve content interoperability with foreign systems, the content metadata can be harvested through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [141].

VP-Core is be a free and open software package. It is based on the Drupal CMS [32] and supports the use of several other Open Source software such as FFmpeg [52].

3.5 Interoperability through Representation Information: OAIS

The concept of *Representation Information* introduced by OAIS [140] provides a very general view of what is needed for understanding and using data. Sub-types of Representation Information include *Structure Information* that imparts meaning about how other information is organized. For example, it maps bit streams to common computer types such as characters, numbers, and pixels and aggregations of those types such as character strings and arrays; *Semantic Information* that further describes the meaning beyond that provided by the Structure Information. It is important to recognize that we must allow a variety of evolving methods of providing these descriptions ranging from text designed for human consumption to complex machine parsable RDF; CASPAR [22] is collecting together a variety of such tools.

3.6 Interoperability through Conceptual Modelling: CSMF

The Conceptual Schema Modelling Facility (CSMF, ISO/IEC-14481 [101]) describes the process of conceptual modelling. The task is to create an abstract formalised description (*conceptual schema*) of some portion of the real world (*universe of discourse*). A *conceptual schema language* provides syntactic and semantic elements used to rigorously describe the conceptual schema in order to convey meaning consistently. A set of principles govern the use of conceptual modelling, including:

- *the 100% principle:* all relevant aspects of a universe of discourse shall be described in the conceptual schema (i.e. a universe of discourse is defined by its conceptual schema)
- *the conceptualisation principle:* a conceptual schema shall contain only aspects that are relevant to a universe of discourse (e.g. it should be independent of physical or technological implementation details)
- *the Helsinki principle:* any meaningful exchange should be based on agreed syntactic and semantic rules, with which a conceptual schema should be formulated and interpreted

Data interoperability is enabled by establishing rules for exchange based on a canonical representation of a conceptual schema.

3.7 Interoperability through distributed systems: RM-ODP

Interoperability for distributed data infrastructures may be developed using the Reference Model for Open Distributed Processing (RM-ODP, ISO/IEC-10746 [82]). Under this model, a system architecture is factored into five complementary viewpoints:

- *Enterprise viewpoint:* purpose, scope and policies governing the activities of the system
- *Information viewpoint:* semantics of information and information processing in the system
- *Computational viewpoint:* a functional decomposition of the system in terms of computational objects and their interfaces (e.g. standardised web service interfaces)
- *Engineering viewpoint:* infrastructure required to support distribution. Whereas the computational viewpoint is concerned with when and why objects interact, the engineering viewpoint is concerned with how they interact (e.g. how to aggregate services at the pan-European level)
- *Technology viewpoint:* specifies particular technology choices for the system

4 Prototypical Use Cases

Interoperability is a tough issue that bridges a multiplicity and diversity of efforts by individual disciplines. The extent of this can be clearly seen from the following few use cases.

4.1 Interoperability of Data Management in the Medical Field

Especially the medical field has already experiences with solutions to data interoperability problems, because of the considerable differences of data models in different medical knowledge domains. In addition, there exists a basic schism in data models and semantics between the field of patient care and the field of clinical research. Perhaps experiences with the creation of interoperability in data management in medicine can set an example for the wider field of interoperability in science. See [138] for an overview of recent approaches to enable interoperability based on the basic idea that in order to achieve technical interoperability, standards have to be applied. In recent decades, many important medical terminologies and vocabularies have been developed. Examples include ICD-10-GM [69], OPS [156], TNM [191], MedDRA [122] and SNOMED CT [173]. But different data and communication standards (HL7 [66] vs. CDISC [12]), different terminology resources (in care coding is done according to ICD9 diagnostic coding [70], CPT coding [21], CPC [20] and a CCS-P [10] coding (surgical coding), Pathology Coding [159], RBRVS Values [165], etc. vs. SNOMED CT: for findings and MedDRA for adverse events in clinical research), different representations of clinical statements (e.g. differences in granularity (systolic blood pressure, measured sitting or standing) and differences in precision (e.g. 65 kg vs. 65,350 kg) exist.

As a high-ranking structure a Reference Information Model (RIM) [168] for Healthcare was established. In this context HL7 v3 offers specifications for data types for health care, XML data formats for medical information, controlled vocabulary and specifications for the Clinical Document Architecture (CDA) [15]. In the domain of clinical research, standards provided by CDISC are used. The mission of CDISC is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of health care. CDISC-based standards cover the following models: Operational Data Model (ODM) [155], Study Data Tabulation Model (SDTM) [180], Analysis Dataset Model (ADaM) [2], Laboratory Data Model (LAB) [111], Protocol Representation Group (PRG) [164], Standards for Exchange of Non-clinical Data (SEND) [175] and Case Report Tabulation Data Definition Specification (CRT-DDS) [9]. Both worlds are bridged by a specific domain model. In general, domain modelling conceptualizes a domain and this conceptualization is represented in computable knowledge as ontologies or domain analysis models. In this way BRIDG focuses on the abstract meaning of concepts shared by clinical research communities. The Biomedical Research Integrated Domain Model (BRIDG) [7], a shared domain analysis model of regulated clinical research, builds a connection with the Reference Information Model (RIM) [168] of Health Level 7 (HL7) [66]. Recently, the CDISC Protocol Representation Model (PRM) [11] that identifies and defines a set of over 300 study protocol elements, was used to map PRM elements to elements of the BRIDG model. This is especially important, because the protocol is the core part of every clinical research study. Thus, PRM protocol information can be readily extracted and entered automatically into information systems or online registries, supporting the general goal of more transparency in clinical trials.

On the practical level, a framework for interoperability between the Electronic Health Record (EHR) [37] and Electronic Data Capture Systems (EDC) [36] is of great importance. For this purpose, several groups have formulated user interoperability requirements and use cases (NHIN Slipstream Project [132], Electronic Health Records/Clinical Research (EHR/CR) project [38], Global EHR/CR Functional Profile Project [56]. OpenEHR [149] is already able to import from messaging systems like HL7 Version 2, for which archetypes have been mapped on HL7 V2 messages. OpenEHR provides a formal model the OpenEHR Reference Model (RM) [148] with *archetypes* as a central element. Archetypes are reusable structured models of clinical information concepts that are used in an EHR. OpenEHR provides proper integration with terminology systems, including SNOMED-CT [173], LOINC [121] and ICD [75].

For clinical research centres the necessity to reflect about the application of harmonized terminologies and metadata dictionaries is based upon four developments: (1) in future, laboratory data, especially genomics data in the context of "personalized medicine", will increase considerably and these data have to be integrated into data management systems; (2) in connection with the European Clinical Research Infrastructure Network (ECRIN) [46] a larger number of international clinical studies will become effective, resulting in the need for multi-lingual systems and extensive trans-national data exchange, (3) the cooperation between medical networks and clinical centres (for example the POH, Oncology and Aids networks of excellence [160]) will require new ways for data integration, (4) the "secondary use" of patient data collected during care sessions for scientific purposes will increase, resulting in new requirements for data harmonisation and interoperability.

Semantic interoperability remains one of the greatest challenges in biomedical informatics. Main aspects of

semantic interoperability cover an information model, common data elements (CDE) [17] and vocabularies (terminologies). Despite the differences in the data models of the different medical domains (e.g. oncology, psychiatry, heart disease), it is to be assumed, that there will be areas, in which standardization and reusability of items and modules are possibly. The need for a comprehensive and flexible model for metadata repositories has been recognised by the Telematics Platform (TMF) [183]. This model should guarantee that metadata models can cover as many illness-related domains and study-related domains as possible. Also possibilities for the representation of complex values, which exist in psychiatric research has to be considered. Demands exists for a semantic standard which connects to terminologies used for data coding (e.g. MedDRA, LOINC and CDISC) and compatibility with important international standards (HL7 RIM, BRIDG, CDISC, ISO-11179 [83]). On the structural level demands exists for synonyms of items, linking of items and a separated storage space of data variations.

To develop metadata repositories the metadata models will have to be based on ISO-11179. The meta-categories of the ISO standard, like data element, data element concept, etc., will have to be adopted in the repository. In the next step, meta-repositories can be supported by the creation of a hierarchy of meta-ontologies, which may build a foundation for shared conceptualizations of knowledge in science.

Any thinking about data interoperability should consider the so-called *Open Source / Open Community / Open Science Approach*. It has been noticed for some time that the increasing complexity of science necessitates new forms of collaboration to enable intensified cooperation, data sharing and networking. Genomics, proteomics and physiological data are increasingly interwoven with clinical and care information. As response to this impeding complexity academia has developed the concept of "open research" described by transparency achieved through open access, open data and open communication. Through open access to data a combined and joint control can be ensured by the members of the network, similar to the common control during code development in Open Source software projects. Free access to information, free communication, mutual control and review by network members, will guarantee transparency of all processes and results, and will require a high level of data interoperability.

4.2 Interoperability in Linguistics

Some aspects of interoperability for the linguistics community, which is part of humanities, involve dealing with a wide variety of resource types, ranging from videos/audios and brain imaging time series to complex structured textual data. The community is faced with interoperability issues at four levels:

- the technical encoding level (UNICODE [193], MPEG1/2/3/4 [125], H.264 [64], linear PCM [120], MP3 [126], etc).
- the structure or format level (XML Schema level for texts, RDF [170] for semantic assertions, etc). The community is working hard on more generic formats for the main data types. These could be called *pivot formats* which would reduce the number of converters that are required to convert the different formats that are currently in use by sub-communities. Some people argue that RDF/XML [170] should be used as the most generic solution for all problems since it just encodes relations of all types in form of simple triples. However, this loses the advantage of structural compactness and constraints which can lead to interpretation problems and inefficiencies.
- the semantic interoperability at linguistic encoding level (concept registries, etc). The linguistics community is collaborating with ISO TC37/SC4 [99] to create a reference registry of linguistic categories (*concepts*). The idea is to give up fixed schemas, but require that schema elements refer to entries in the data category registry. In doing so the community can start solving the semantic interoperability problem where possible. Yet it is far away from being able to claim that different tag sets that emerged over decades of linguistic theory can be mapped easily on each other.
- the content interoperability level (domain ontologies, etc).

At the first level much has been done and except for migrations to new formats things have settled. What the community is currently looking at intensively are the second and third levels. At the second level it is defining generic pivot formats (generic models with some instantiations such as XML schema based) and making them ISO [77] standards - the first standard of this sort is Lexical Markup Framework [117], just standardized by ISO.

At the third level the community has just defined a standard called ISO-12620 [84] where it has defined a model for specifying concepts (due to their reduced definition ISO-12620 prefers to speak about data categories) at sufficient detail. These registered concepts can be seen as point of references to bridge between different phenomena encoding schemes.

4.3 Interoperability in an open e-Humanities ecosystem

The e-Humanities are an immensely diverse and dynamic community. Apart from a few sub-fields such as the psycholinguistics (cf. respective use case), there are few authoritative standards on a content level that a whole (sub-)community complies with. Humanists hold their freedom high to express and encode without constraints. Acknowledged standards just like the XML-based Text Encoding Initiative (TEI) [184] are themselves masterpieces in flexibility - two TEI-annotated texts can be, but are not necessarily interoperable to a given degree. TEI is very flexible, so tags are embedded in rather different contexts making interpretations a very hard job. When researchers are using TEI they use different flavours. Some just make use of some tags, but use a completely different schema framework to overcome the complexity.

The TextGrid project [185], which is among the spearheading initiatives establishing e-Humanities infrastructure, established a layered approach to achieve interoperability in this open environment. The prime objectives for its approach were in low entry barriers, and nurturing opportunities rather than enforcement (the carrot rather than the stick).

While TextGrid innovators are domain researchers from the philologies and linguistics with their own ideas about how to encode domain semantics "the right way", it is their conviction that rigid guidelines hamper participation more than they enable interoperability. Eventually an interoperable environment without any data fails to be useful. TextGrid thus puts enabling participation over interoperability, while fostering interoperability wherever it can.

The approach is simple: the user can do whatever she wants, but by offering interoperable data according to TextGrid recommendations she increases exposure and is provided with more functionality in the TextGrid virtual research environment. As such, the layers can be sketched like this:

- any data format can be uploaded, TextGrid ensures bit-preservation
- metadata facilitates data management and retrieval (metadata-based search)
- by uploading XML-based texts, a series of services can be used on the data including streaming tools, an XML-editor, and other functionalities
- if the XML follows TEI encoding, TextGrid offers graphical editing, metadata extraction, and other functionalities
- defining a mapping to the TextGrid recommendation for a TEI core encoding allows interoperability on a semantic level

In its design of the incentive system, TextGrid follows basic principles of collaborative environments in the Web 2.0 world today. Moreover, it is designed to be open, and is expected to grow over time. There may even be competing incentive systems within TextGrid at some point in the future.

This use case illustrates an open, participatory approach to achieving interoperability. It is not the role of e-Infrastructure to judge how research is done correctly. The art of building e-Infrastructure really lies in enabling the users to achieve their goals through intuitive and simple user interfaces that integrate with their daily work environment. Network effects of collaborative environments and incentive systems are powerful mechanisms that enable interoperability to emerge.

4.4 Interoperability in Earth Sciences

One of the most fundamental challenges facing humanity at the beginning of the 21st century is to respond effectively to the global changes that are putting increasing pressure on the environment and on human society. Climate

change, biodiversity and habitat loss, ocean acidification, environmental degradation and sustainability are among the aspects that need to be understood and managed. By their very nature, these problems extend beyond political and administrative boundaries, and demand a policy response grounded firmly in a robust evidence base. For example, the United Nations Framework Convention on Climate Change (UNFCCC) [197] and its Kyoto protocol were negotiated on the basis of the data-based analyses of the Intergovernmental Panel on Climate Change (IPCC). The latest IPCC report [79] changed from "likely" to "very likely" the assessment in its previous report that recent observed warming is due to an increase in anthropogenic greenhouse gas concentrations – a change that carried enormous political implications [131]. It also noted the crucial role of data in this revised assessment (emphasis added):

- "Scientific progress since the Third Assessment Report (TAR) is based upon *large amounts of new and more comprehensive data, more sophisticated analyses of data*, improvements in understanding of processes and their simulation in models and more extensive exploration of uncertainty ranges." (IPCC, 2007: Summary for Policymakers. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [174][79])

The demand for access to continuously increasing quantities of heterogeneous data and resources (mechanisms, infrastructure, tools, etc.) needed to achieve such objective is growing. The need to face this increasing demand highlights two basic aspects on which the scientific community should focus:

- Data, coming from heterogeneous sources as needed in the different domains;
- Resources, to be used to exploit such data in an effective way.

Unfortunately, for policymakers – as well as scientists and the general public – locating and accessing the right data, products, resources and other information needed to turn data into knowledge is not always easy. At each stage of the pipeline (discovery, access, use), a legacy of non-interoperable data and metadata formats and services impedes the information flow.

Today, indeed, information about the state of the Earth, relevant services, data availability, project results and applications are accessible only in a very scattered way through different operators, scientific institutes, service companies, data catalogues, etc. Referring to remote sensing missions, only a limited community with specific knowledge of what to search for, is today in a position to collect, compile and thus exploit the necessary Earth Observation (EO) information.

It is as well not to be forgotten that each Earth Science community domain may need specific methods, approaches and working practices for gathering, storing and exchanging data and information.

This demand of the ES community shows the need for an efficient data infrastructure able to provide reliable, easy, long-term access to Earth Science data via the Internet, so to allow Earth scientists and users to easily and quickly derive objective information and share knowledge based on all environmentally sensitive domains.

Interoperability appears to be a key issue in the development of an efficient and value-adding data infrastructure.

Interoperability efforts for data sharing Significant efforts are being made to develop an interoperability framework that will enable unprecedented environmental data sharing and integration opportunities. For example, the recent INSPIRE Directive (2007/2/EC) [72] requires all European public authorities holding 'spatial data' [1] to provide access to that data through common metadata, data and network service standards. Crucially, it does not require data custodians to refactor their existing databases (though they may choose to do so), only to provide an interoperable, standards-based 'view' onto that data. INSPIRE implementation costs have been estimated [44] at between 100-300 million euro per year for ten years, but with annual net benefits of over 1000 million euro.

[1] The Directive covers thematic data in any of 34 areas: Coordinate reference systems, Geographical grid systems, Geographical names, Administrative units, Addresses, Cadastral parcels, Transport networks, Hydrography, Protected sites, Elevation, Land cover, Orthoimagery, Geology, Statistical units, Buildings, Soil, Land use, Human health and safety, Utility and governmental services, Environmental monitoring facilities, Production and industrial facilities, Agricultural and aquaculture facilities, Population distribution – demography, Area management/restriction/regulation zones and reporting units, Natural risk zones, Atmospheric conditions, Meteorological geographical features, Oceanographic geographical features, Sea regions, Bio-geographical regions, Habitats and biotopes, Species distributions, Energy resources, Mineral resources.

The interoperability framework for spatial and environmental data within INSPIRE is being developed primarily through ISO Technical Committee 211 (TC211) [100] on Geographic information and Geomatics; conformant infrastructures (e.g. INSPIRE) are known as 'Spatial Data Infrastructures' (SDIs) [57]. ISO TC211 is developing a suite of *de jure* standards (currently numbering over 50 issued or under development) known by their project numbers as the 'ISO 191xx series'. In addition, the *de facto* non-profit standards body, Open Geospatial Consortium (OGC) [143], runs accelerated specification development activities and enjoys a productive 'Class A liaison' with ISO TC211.

Moreover, standards-based interoperability for geospatial and environmental data may provide a template for cross-DG action in other areas – in Europe, standards have enabled regulatory intervention (INSPIRE) by EUROSTAT, DG-ENV, and the JRC (see INSPIRE open letter [73]), backed up by various DG-INFSO FP7 funding actions.

The basis for interoperability in SDIs is defined on two levels – the conceptual level, and the implementation level – both of which draw on existing foundational models.

At the conceptual level, the Conceptual Schema Modelling Facility (see section 3.6) is applied by the ISO TC211 series of standards to develop conceptual schemas for datasets. The Unified Modelling Language (UML) [194] is adopted as the CSMF conceptual schema language. The 'General Feature Model' (ISO-19109) is also defined as a general (object-like) meta-model: a 'feature' is an abstraction of a real-world phenomenon (e.g. 'bridge'), and may be characterised by its attributes (e.g. 'height', 'width'), relationships with other features (e.g. 'crosses road'), and operations that may be performed (e.g. 'widened'). Feature types may inherit from one another, and may be catalogued for reference and re-use in a *Feature Catalogue* (ISO-19110) [90].

With the General Feature Model as meta-model, the ISO TC211 'Domain Reference Model' (ISO-19101) [101] then provides an interoperability framework for data. A *Dataset* contains features (including attributes, relationships with other features, operations). The logical structure and semantic content of a Dataset is described through an *Application Schema* (expressed using the UML conceptual schema language, and using where needed other standardised conceptual building blocks [2] and feature types). A *Metadata Dataset* allows search and evaluation of data, and is itself structured according to a standardised conceptual schema (ISO-19115 [94][95]), which may include references to the dataset's defining application schema, feature catalogues etc. Finally, standardised network *Services* provide access and processing (e.g. search, transformation) functionality on both metadata and data. Examples of standardised web services include a viewing service (ISO-19128 [93]) for rendering selected data as an image, and a data access service (ISO-19142 [96]) for retrieving selected data, in a standardised structure and format, from a remote database.

At the implementation level, Spatial Data Infrastructures are architected for interoperability as distributed systems using the Reference Model for Open Distributed Processing (see section 3.7). The General Feature Model and metadata schema are important components of the Information Viewpoint, while standardised web service interfaces for data discovery, view and download provide the Computational Viewpoint. Data sharing/licensing agreements are part of the Enterprise Viewpoint.

How interoperability is addressed by GENESI-DR: *GENESI-DR* (Ground European Network for Earth Science Interoperations - Digital Repositories) [63] is an ESA-led [47], European Commission (EC)-funded two-year project, that is taking the lead in providing reliable, easy, long-term access to Earth Science data. GENESI-DR allows scientists from different Earth Science disciplines located across Europe to locate, access, combine and integrate Earth-related data from space, airborne and in-situ sensors archived in large distributed repositories. A dedicated infrastructure providing transparent access to all this will support Earth Science communities by allowing them to easily and quickly derive objective information and share knowledge based on all environmentally sensitive domains.

GENESI-DR is a response to most of the Earth scientists needs for interoperability. The *Central Discovery Service* allows to query information about data existing in heterogeneous catalogues, at federated DR sites in a transparent way and can be accessed by users via web interface, the *GENESI-DR Web Portal*, or by external applications via open standardized interfaces (OpenSearch-based [154]) exposed by the system.

[2] Many of the ISO-191xx standards provide standardised conceptual models for, e.g., spatial geometry (ISO-19107 [91]), time (ISO-19108 [88]), coordinate reference systems (ISO-19111 [89]), physical 'fields' (e.g. temperature distributed over time and/or space, ISO-19123 [92]), etc.

Attention to standardisation in modelling information, flexibility and scalability of the architecture allow easy integration of new Digital Repositories. GENESI-DR information objects follow the ISO-19115 [94][93] standard for describing geographic information and services (and the corresponding XML schema implementation ISO-19139 [97]) and the INSPIRE [71] Implementing Rules to use ISO-15836 (Dublin Core) [86]. Together they provide a profile that fits very closely with INSPIRE guidelines but allows more precision and models metadata about geographic information where the usage of Dublin Core is maximized.

To share information about datasets and collection available on the Digital Repository nodes the GENESI-DR information model is defined as an RDF model rendered in XML [170].

Different and efficient data transfer technologies such as HTTPS [67], GridFTP [134] and BitTorrent [8] are used to guarantee easiness and fast access to large volumes of distributed data; GENESI-DR foresees services to enable expert users exploiting computational and network resources in order to produce the final desired product. This means either passing input data to a processing service (an OGC [144] Web Processing Service, for instance) available at some site or running a user application/algorithm on Grid resources on specified data sets.

Another objective of GENESI-DR is the adoption of a data curation and preservation strategy in order to preserve the historical archives and the ability to access the derived user information as both software and hardware transformations occur.

In this context the OAIS model [140] and the PREMIS vocabulary [162] (for preservation metadata) are analysed.

GENESI-DR, being a data centric and interoperability based e-infrastructure, can represent the most comprehensive solution to the ES needs: interoperability, data heterogeneity management and multi-disciplinarity requirements are in fact met and considered in the appropriate direction. At the same time inside this framework enhancements are desirable in order to support and satisfy the dynamically emerging ES needs to maximize interoperability, operativity and inter-disciplinary collaboration.

4.5 Interoperability in Astronomy and Space Science

Astronomy faces a data avalanche [158][1]. Breakthroughs in telescope, detector, and computer technology allow astronomical instruments to produce terabytes of images, catalogues, spectra and time domain products. These datasets may cover the entire sky in different wavebands, from gamma- and X-rays, optical, infrared, sub-mm through to radio. The combination of inexpensive storage technologies and the availability of high speed networks has led to the concept of multi-terabyte online databases interoperating seamlessly. This is known as the Virtual Observatory (VObs). More and more data sets are becoming interlinked via data publishing efforts such as European FP6 and FP7 programs coordinated by the European Virtual Observatory (Euro VObs) [187] and query engines are becoming more and more sophisticated [190]. Moore's law is driving astronomy even further: new survey telescopes are being constructed which will image the entire sky every few days and yield data volumes measured in petabytes. These technological developments are driving fundamental change in the way astronomy is done. For example, the detection of new events in transient objects can automatically trigger the coordinated follow-up at other facilities including robotic telescopes within seconds of initial discovery by use of the VOEvent protocol [201].

Astronomers and space scientists set up the International VObs Alliance (IVOA)[78], which was formed in June 2002 with a mission to "facilitate the international coordination and collaboration necessary for the development and deployment of the tools, systems and organizational structures necessary to enable the international utilization of astronomical archives as an integrated and interoperating virtual observatory." The IVOA comprises many national VObs projects including Armenia, Australia, Canada, China, Europe, France, Germany, Hungary, India, Italy, Japan, Korea, Russia, Spain, the United Kingdom, and the United States as well as international organisations such as the European Space Agency (ESA) and the European Southern Observatory (ESO). Membership is open to other national and international projects according to the IVOA Guidelines for Participation [104], and Brazil has recently joined.

The work of the IVOA focuses on the development of standards. Working Groups are constituted with cross-project membership in those areas where key interoperability standards and technologies have to be defined and agreed upon. The Working Groups develop standards using a process modelled on the World Wide Web Consortium, in which Working Drafts progress to Proposed Recommendations and finally to Recommendations [103].

Recommendations are ultimately endorsed by the Virtual Observatory Working Group of Commission 5 (Astronomical Data) of the International Astronomical Union [188]. Current working groups include Applications, Data Access Layer, Data Modelling, Grid and web services, resource Registry, Semantics, VOEvent, VOQuerylanguage, VOTable. The IVOA also has Interest Groups that discuss experiences using VObs technologies and provide feedback to the Working Groups covering the areas of theory, OGF and data curation and preservation. Reports on progress and future roadmaps are issued annually [189].

Senior representatives from each national VObs project form the IVOA Executive Committee. A chair is chosen from among the representatives and serves a one-year term, preceded by a one-year term as deputy chair. The Executive Committee meets 3-4 times a year to discuss goals, priorities, and strategies. The IVOA holds two Interoperability Workshops each year: a week-long meeting in spring and a shorter meeting in fall that is either coordinated with the annual Astronomical Data Analysis Software and Systems conference or with a regional VObs project meeting. These meetings are opportunities for the Working Groups and Interest Groups to have face-to-face discussions and for the more difficult technical questions to be resolved.

The recently completed EU FP6 EuroVO Data Centre Alliance project [187] has conducted a first ever census of European data centres in astronomy and space science which will form the basis of increased coordinated efforts in the coming years in the area of standardised astronomical data publishing. European participation in the development of IVOA interoperability standards is coordinated by the FP7 Euro-VO Astronomical Infrastructure for Data Access (EuroVO-AIDA) project.

Data providers publish their data sets in the VObs by implementing a VObs-compliant *interoperability layer* on top of their data holdings. Services access the wealth of distributed, heterogeneous data, and communicate with each other, using VObs protocols. European VObs projects provide several key services, including portals to image and spectral data such as Aladin and VOSpec, provided respectively by the French VObs and European Space Agency VObs, and the Astrogrid VODesktop, the most complete end-to-end solution to date for accessing, filtering, visualising and manipulating VObs datasets, which was made available by the UK VObs Technology Centre project [186] in 2008.

The semantic technologies which have been deployed so far within the IVOA range from very lightweight data description vocabularies to intricate ontologies which can be used for database consistency checks. The IVOA's Unified Content Descriptors (UCDs) [198], comprise a very pragmatic vocabulary for data, which was derived from the existing database schemas at a current large archive. Although they do not solve the general problem, these have been effective in promoting real interoperability between multiple archives and applications. At a slightly higher level, there is also an IVOA standard for Vocabularies [106] which mandates the use of the W3C SKOS standard for developing controlled vocabularies and thesauri. Additionally, the IVOA is developing an Ontology of Astronomical Object Types [105] as an OWL-1.1 ontology, and in doing so is exploring the possibilities of very sophisticated reasoning to validate database entries.

4.6 Interoperability in Particle Physics

By far the largest particle physics project is the Large Hadron Collider (LHC) [113], with its associated computing infrastructure, the LHC Computing Grid (LCG) [119]. It would be inaccurate to see this as a single community:

- The tiered data management model (one Tier 0, about 11 Tier 1s, many Tier 2s and Tier 3s) means resource providers are very diverse and no single data storage implementation will fit all;
- Each LHC experiment (Alice, Atlas, CMS, LHCb) has its own model for moving and analysing data between tiers and between sites in the same tier;
- Resources are also usually being used not only by non-LHC particle physics resources but also often by non-physics communities;
- In particular, service providers to LCG currently include OpenScienceGrid (OSG) [153] in the US and EGEE [40] in Europe. These Grids are also being used by other communities.

Consequently, there is a strong need for resource interoperability in data management in LCG. The SRM implementations mentioned in Section 2 are an example of this: interoperability was originally mainly driven by the requirements of LCG.

Of course, there are other particle physics experiments. As individual scientists sometimes participate in more than one experiment, and service providers certainly provide services to more than one, we see again the need for interoperability: scientists are interested in using their existing applications and software libraries for data analysis, and service providers are interested in supporting fewer interfaces.

5 Miscellaneous Related Issues

5.1 Issues with Models

Existing data of all types and provenances is for historical and other reasons scattered across systems constructed according to *data models* that are either hierarchical, relational, XML or object based. Peer-to-peer federation is generally not possible, and client-server aggregation is not easy, across heterogeneous models. In general, interoperability across data models is likely to be easier to achieve if some form of translation is employed, either directly or via a neutral format.

Regarding *metadata*, the term is used to mean "data about data" generally or specifically "data describing resources". A general method of modelling information is the Resource Description Framework (RDF) [170], a family of W3C specifications based upon the idea of making statements about Web resources in the form of subject-predicate-object expressions (called *triples* in RDF terminology). The subject denotes the resource, and the predicate denotes aspects of the resource and expresses a relationship between the subject and the object. RDF is an abstract model with several possible serialization formats (i.e., file formats, such as XML). The way in which a resource or triple is encoded depends on the particular format. RDF is therefore a foundation for processing metadata and it provides interoperability between applications that exchange machine-understandable information.

RDF metadata can be used in a variety of application areas; for example: in resource discovery to provide better search engine capabilities; in cataloguing for describing the content and content relationships; by intelligent software agents to facilitate knowledge sharing and exchange; etc. The resources being described by RDF are, in general, anything that can be named via a URI (Uniform Resource Identifier). The broad goal of RDF is to define a mechanism for describing resources that makes no assumptions about a particular application domain, nor defines the semantics of any application domain.

RDF in itself does not contain any predefined vocabularies for authoring metadata. Anyone can design a new vocabulary, the only requirement for using it is that a designating URI is included in the metadata instances using this vocabulary. This use of URIs to name vocabularies is an important design feature of RDF: many previous metadata standardization efforts in other areas have foundered on the issue of establishing a central attribute registry. RDF permits a central registry but does not require one. For example, an RDF model rendered in XML is the basis of the GENESI-DR [63] information model to share information about datasets and collections available on the Digital Repository nodes.

However the interpretation tends to be very discipline specific; it is then very hard to be sure that one has all the metadata that is needed. For example very often metadata is taken to mean essentially that which is needed to discover the material of interest. However discovery is only the first step which is required. The OAIS Reference Model (ISO-14721) [140][85] provides a finer taxonomy of metadata. Types of metadata information include: *Representation Information* that is needed interpret binary object into something that can be understood; *Descriptive Information* that is needed for discovery; *Preservation Description Information* that is needed to verify authenticity. Each of these has further sub-divisions. The CASPAR project [22] has implemented many of the fundamental components, in particular those for Representation Information, including virtualisation techniques, which facilitate interoperability between systems.

Over time many things change including hardware, software and the knowledgebase of users. The OAIS concept of Representation Information is itself a piece of information, which needs its own Representation Information. This recursive concept defines a *Representation Network* which together provides enough for a designated community to understand the data of primary interest. This changes over time and therefore the Representation Network must be updated over time to ensure the usability and interoperability of data. *Registries of Representation Information* could provide an important way of sharing and enhancing Representation Networks.

CLARIN is working on generic data models which they call *meta-models*. These are different from metadata models, since metadata in the restricted sense means "keyword type of resource descriptions such as Dublin Core".

Here CLARIN is working on a flexible meta model for component metadata which is called the CLARIN Metadata Initiative (CMDI) [14].

Interoperability versus Time: Data and model formats can only be expected to have a finite lifetime (typically short), and so those that maintain data repositories face the question of whether to (a) continue to provide *compatibility* for an extinct format indefinitely or (b) to provide *translation* for any dependent material either directly of via a neutral format.

The former is true *interoperability*, driven by a desire that all data must be reachable and usable by all. This may not scale. The pure visibility and accessibility of data does not per se lead to scalability problems. Catalogues may become huge if they harvest all metadata descriptions that are available, but this will not occur for reasons of domain scope. Why should one, for example, harvest all resources from astronomy and plasmaphysics. While there is a challenge in how to present the huge mass of data, how to allow selections, filtering, etc, the lack of scalability is instead in terms of effort and cost required to continuously provide, maintain and update such universal compatibility.

The latter is an attempt to achieve *scalable interoperability*. Short-term fixes adopting this approach yield *interoperation*. Long-term frameworks that achieve this provide true *interoperability*.

Scalable Interoperability: One may predicate that interoperability does not necessarily mean all data must be reachable and usable by all. This requires some form of (possibly lossy) data workflow to translate between source and destination environments. It is more achievable if there is:

- no universality unless the need is proven; the effort and cost of handling rare combinations is deferred until the proven need arises (*lazy implementation*).
- no infrastructural bias (*neutral exchange*). A bad example of infrastructural bias might be data exchange via MPI-1, with its limited typing. Neutral exchanges have the advantage of being agnostic regarding the data representations they handle. DTD/RDF/XML is possibly the best developed example.

However, data should be widely accessible by design, and even if a neutral exchange is not employed, mechanisms should be provided to facilitate the handling of corner cases if and when the need arises.

Interoperability and Workflows: There are various ways that data workflows can be used to benefit interoperability. Some repository systems allow the definition of internal data workflows, for example the *iRODS* [80] system allows data workflows to be defined within *rules* that can be built up cumulatively from other rules, with data passed into/within rules; a rule is defined as `actionDef | condition | workflowChain | recoveryChain` (*actionDef* identifies the action to be carried out; *condition* must be TRUE for execution; *workflowChain* is a sequence of actions to be executed; and *recoveryChain* is a sequence of recovery actions to ensure consistent state after an error). Alternatively a generic workflow engine such as *Kepler* [203], *WebCom-G* [109] or *Taverna* [182] can be used to establish a chain of actions to provide interoperability. In astronomy, the UK VO Technology Centre Astrogrid project [186] provides a higher level programming language interface using python scripting for the more complex multi archive and multi tool queries [204] while a modular workflow solution using *Taverna* will be included in a 2009 release.

Workflow interoperability then becomes important, e.g. to allow re-use of a workflow defined for one workflow engine on another, or workflow cross-invocations in a heterogeneous workflow infrastructure. Fundamental aspects that affect interoperability between workflow systems are discussed in [39].

5.2 Issues with Supporting Mechanisms and Policies:

Data management is increasingly becoming policy-driven. There are issues like curation/privacy/liability/etc that will need serious consideration and are likely to drive policy, and hence interoperability must consider the propagation of policy.

Governance of Distributed Heterogeneous Facilities Governance of distributed heterogeneous facilities is different from governance of large homogeneous facilities. It is very important to establish good governance structures for distributed heterogeneous data structures and flexible (international) cooperation agreements to facilitate access to such data.

Interoperability of Data Access Policy Management Mechanisms: Both institutional and federated repositories will require access control. This involves many aspects, for example, data access authorization should be uniform from a user's viewpoint at any location; it should not happen that a user is authorised to access a replica of the data at one location but not on another. Authorization should be done on a VO level; this has been demonstrated by operational experience at many levels in many large and small collaborations. Users must be able to access data locally outside of the data management context (i.e. to have *backdoor access*), but this must also be able to be denied. An institution must be able to allow/deny access by specific users to resources at that institution, and must be able to audit usage including the name of the user.

It is clear from widespread previous experience that a spectrum of desired access rights will arise, from anonymous, to relatively open access to those for whom identity can be established via federated identity management [50] (e.g. OpenID [150]), to stronger controls based on grid-like PKI [62], which must subsume or extend to the IGTF global PKI infrastructure [76]. For most scientific communities, the mechanisms for achieving this are not yet decided at the national level, let alone internationally.

Access controls primarily involve authentication, auditability, authorization, user and group access controls, possibly role-based access controls (RBAC), and VO access controls. There are many possible combinations of available methodologies. For example, a community might propose that:

- Authentication and authorization events should be recorded in an auditable fashion (if possible, with traceability of an access to its ultimate source action).
- For anonymous access the authorization should be able to be conditioned by policies.
- For a successful user access by an individual, the authorization should be able to be conditioned according to their identity, the group they belong to, the role they are acting in, the VO they are a member of, and policies.
- For a successful access by a member of a group, the authorization should be able to be conditioned according to the group they belong to, the role they are acting in and the VO they are a member of, and policies. Such an access might be used by an editorial group of a digital repository.
- For a successful access by a member of a VO, the authorization should be able to be conditioned according to the role they are acting in and the VO they are a member of, and policies. Such an access might be used by pilot (*placeholder*) jobs that are submitted by specific authorized members of a VO.

None of the above prescribes the form of a credential, where the identity is authenticated, where access control information resides, where it is used to make authorization decisions, or where these decisions are enforced, but whatever the choices the community makes, the data will not be widely accessible unless uniform or interoperable mechanisms are employed.

Interoperability of Data Privacy Environments: Who can access the data? Privacy often mandates authorization, anonymization and encryption facilities, and in addition often requires policies that preclude un-encrypting the data in facilities that are not authorized to do so; all this may require additional trust-based authorization mechanisms and policies. IPR and data protection aspects of privacy policy will have (possibly international) legal ramifications. Privacy policy must include disclosure to law-enforcement agencies and data protection as mandated by legislation.

Interoperability of Legal Environments: Similarly where data is federated or strong dependencies exist, who is responsible in the case of data loss? Liability, applicable law, arbitration, etc, issues may initially be deferred by provision of a standardised waiver of liability in a VO's *Acceptable Usage Policy (AUP)*, and for example Grid-Ireland proposes to deploy continuous automatic checking of VO AUP's for the existence of the standardised

waiver. Such solutions will not, however, suffice for the long-term. Access rights and management may also vary from one legal environment to another and the rights themselves may change over time, and this needs to be modeled (for example, CASPAR [22] is developing a high level view of rights management and legal frameworks which will allow one to derive rights over time and between legal frameworks). The e-IRG Support Programme (e-IRGSP2) has conducted a large-scale study of the legal issues related to e-Infrastructures, especially across jurisdictions, many of which relate to data [116].

Interoperability of Data Snapshot/Archive Policy Management Mechanisms: *Snapshots* are a convenient means of capturing the current state of a data repository, i.e. *live archival*, but in general can only be restored to an identical system architecture. Virtualization is likely to prove valuable in exchanging snapshots, from which other mechanisms can then extract data of interest.

Regarding *archives*, a reference model for the design of an archive, that also focusses on the necessary concepts and ways to proceed for long term digital information preservation and access, is the "Reference Model for an Open Archival Information System" (OAIS) (ISO-14721:2003 [140][85]). OAIS distinguishes between an Information Package that is preserved (*Archival Information Package* or AIP) and the Information Packages that are submitted to the OAIS by the Producers (*Submission Information Packages* or SIP). The Archival Information Package contains *Content Information* (the target of preservation) and *Preservation Description Information*. CI and PDI shall be packaged and equipped with *Packaging Information*. Descriptive information (the set of information necessary to support the finding, ordering, and retrieving of data) shall be associated to each package. The CASPAR [22] components provide a basis to produce an AIP complying with the OAIS directives starting from a generic SIP.

Interoperability of Data Curation Policy Management Mechanisms: Curation policies typically provide for selective migration of data to secondary stores and finally to physical media (and then often only if strictly necessary). Curation of the data will also be necessary for the retention of unique observational information which is impossible to recreate (*Ulieta Bird effect*), the retention of expensively generated data which is cheaper to maintain than to recreate, the reuse of data for new or future research purpose, in order to validate and account for publicly funded research, and for compliance with legal requirements for educational and teaching purposes. Where data is federated or strong dependencies exist, interoperable curation policies are desirable to guarantee the ability to completely restore the data to the satisfaction of all federated and dependent parties.

Retention of Policy Management Information: For all such policies, is policy information retained? If it is held externally, is it imported into snapshots/archives/curation? Again, if it is held externally, is the necessary information retained? Is the provenance of policy information retained? Are the minimum requirements for legal proofs retained ?

6 RECOMMENDATIONS

- Actively encourage programmes that support cross-disciplinary access to digital objects and related services.
- Encourage the development of non-discipline-specific frameworks and information architectures for interoperable exchange of data;
- Support communities for the definition of their requirements and their activities in the domain of semantic interoperability;
- Support interoperation activities within multinational **and** multi-disciplinary/community grids, e.g. OGF activities, or within EGI; the activity itself, however, is likely to be focused on a part of the infrastructure, e.g. authentication, job submission, or storage;
- Prioritise those interoperation activities aiming at standardising interfaces and/or protocols, or documenting current usage and limitations of existing standards for interfaces and protocols;
- Ensure that work is practical and realistic instead of theoretical "paperwork";

- Ensure that besides hardware and services, digital objects deserve infrastructure components in their own right:
  - mediation services for metadata / semantic annotations of data;
  - persistent linkage of research data with publications and other resources;
  - policies for long-term preservation of data, maybe focused into dedicated centres (preservation activities plus consultation);
- Define proper governance structures and guidelines for (inter)national agreements for distributed heterogeneous data facilities;
- Highlight that the basis of proper data management is a proper repository set up with strict organizational guidelines that are supported as widely as possible by a proper repository system;
- Highlight that for achieving semantic interoperability in open scenarios the project oriented approach of formal ontologies seems to be problematic, suggesting that a separation between concept definitions and their relations is desirable (as is suggested by ISO 11179 and ISO 12620 for example);
- Highlight that state-of-the-art network infrastructures are needed that are capable of adapting flexibly to the needs of the applications and researchers relying on them.

7 DMTF-INTEROP MEMBERSHIP

Prenom	Nom	Country	Institution	eMail
Patrick	Aerts	NL	NWO	aerts at nwo.nl
Andreas	Aschenbrenner	DE	University of Gottingen	aschenbrenner at sub.uni-goettingen.de
Hilary	Beedham	UK	UK Data Archive, University of Essex	beedh at essex.ac.uk
Brian	Coghlan	IE	Computer Science, Trinity College Dublin	coghlan at cs.tcd.ie
Rudolf	Dimper	EIRO	ESRF Computing Services Division	dimper at esrf.fr
Luigi	Fusco	EIRO	European Space Agency	Luigi.Fusco at esa.int
Françoise	Genova	FR	Strasbourg Astronomical Data Centre	genova at newb6.u-strasbg.fr
David	Giaretta	UK	STFC	david.giaretta at stfc.ac.uk
Maria	Koutrokoi	GR	Ministry of Development, Athens	mkoutr at gsrt.gr
Diego	Lopez	ES	RedIRIS, Spain	diego.lopez at rediris.es
Christian	Ohmann	DE	Heinrich Heine University, Duesseldorf	Christian.Ohmann at uni-duesseldorf.de
Wolfgang	Kuchinke	DE	Heinrich Heine University, Duesseldorf	kuchinkw at uni-duesseldorf.de
Pasquale	Pagano	IT	ISTI, CNR	pasquale.pagano at isti.cnr.it
Miroslav	Tuma	CZ	Academy of Sciences, Czech Republic	tuma at cs.cas.cz
Dany	Vandromme	FR	Renater, Paris	dany.vandromme at renater.fr
Peter	Wittenburg	NL	Psycholinguistics, Max Planck, Nijmegen	peter.wittenburg at mpi.nl
Andrew	Woolf	UK	STFC	andrew.woolf at stfc.ac.uk
Hans	Zandbelt	NL	SURFnet, Amsterdam	hans.zandbelt at surfnet.nl
Matti	Heikkurinen	CN	Emergence Tech Ltd.	matti at emergence-tech.co.uk
Michèle	Landes	FR	Renater, Paris	landes at renater.fr
David	Corney	UK	STFC, Didcot	david.corney at stfc.ac.uk
Jonathan	Tedds	UK	Univ.Leicester	jat at star.le.ac.uk
Jens	Jensen	UK	STFC, Didcot	jens.jensen at stfc.ac.uk
David	O'Callaghan	IE	Computer Science, Trinity College Dublin	david.ocallaghan at cs.tcd.ie

Table 3.1: DMTF-INTEROP Membership

8 Definitions, Acronyms and Abbreviations

API Application Programming Interface

DMTF e-IRG Data Management Task Force

DMTF-INTEROP e-IRG Data Management Task Force Interoperability Subgroup

e-IRG e-Infrastructures Reflection Group, see [35]

ESFRI European Strategy Forum on Research Infrastructures, see [48]

EU European Union

GSI Globus Security Infrastructure, see [62]

NREN Research and education network provider

VO Virtual Organisation

VObs Virtual Observatory

IVOA International Virtual Observatory Alliance

9 References

[1] A Discussion Paper for the OECD Global Science Forum: http://www.ivoa.net/pub/info/OECD-QLH-Final.pdf

[2] Analysis Dataset Model (ADaM): available from CDISC website: http://www.cdisc.org/models/adam/V2.0/index.html [cited 2009 May 12]

[3] ARCA: http://eunis.dk/papers/p79.pdf and also: http://www.terena.org/activities/media/ws1/slides/d2-2-ARCA-VRRD.pdf

[4] ARDA Metadata Catalogue (AMGA): http://amga.web.cern.ch/amga/

[5] Baud, J.P., Casey, J., Lemaitre, S., Nicholson, C., Smith, D., and Stewart, G., *LCG Data Management: From EDG to EGEE*, UK e-Science All Hands Meeting, Nottingham, 2005: http://ppewww.physics.gla.ac.uk/preprints/2005/06/

[6] Beckett, G., *Building a Scientific Data Grid with DiGS*, UK e-Science All Hands Meeting, September 8-11, 2008: http://www.allhands.org.uk/2008/programme/download.cfm?id=1056&p=pdf

[7] Biomedical Research Integrated Domain Model (BRIDG): http://www.bridgmodel.org/

[8] BitTorrent: http://www.bittorrent.com/

[9] Case Report Tabulation Data Definition Specification (CRT-DDS): available from CDISC website: http://www.cdisc.org/models/def/v1.0/index.html [cited 2009 May 12]

[10] CCS-P coding (surgical coding): Schnering, P., *Professional Review Guide for the CCS-P Examination*, Delmar Cengage Learning, 2007

[11] CDISC Protocol Representation Model (PRM): http://www.cdisc.org/standards/protocol.html

[12] CDISC: available from Clinical Data Interchange Standard Consortium: http://www.cdisc.org/ [cited 2009 May 12]

[13] Certification of Digital Archives Project: http://www.crl.edu/content.asp?l1=13&l2=58&l3=142

[14] Clarin MetaData Initiative (CMDI): http://www.clarin.eu/files/wg2-4-metadata-doc-v5.pdf

[15] Clinical Document Architecture (CDA): available from HL7 website: http://www.hl7.org.uk/version3group/cda.asp [cited 2009 May 12]

[16] Cohen, D., *On Holy Wars and a Plea for Peace*, IEEE COMPUTER, pp.48-54, October 1981.

[17] Common Data Elements (CDE): http://apiii.upmc.edu/pdf/E%20Poster%20Abstracts%202%2020071.pdf

[18] Common Language Resources and Technology Infrastructure (CLARIN): http://www.clarin.eu/

[19] COPA: http://www.rediris.es/ldap/copa/copa-intro.en.pdf

[20] CPC: available from American Academy of Professional Coders website: http://www.aapc.com/certification/cpc.aspx [cited 2009 May 12]

[21] CPT coding: available from American Medical Association website: https://catalog.ama-assn.org/Catalog/cpt/cpt_search.jsp [cited 2009 May 12]

[22] Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR): http://www.casparpreserves.eu/

[23] d-Cache: http://www.dcache.org/

[24] Data Documentation Initiative (DDI): http://www.ddialliance.org/codebook/

[25] Data Movement Interface (OGSA-DMI): http://www.ogf.org/gf/group_info/charter.php?review&group=ogsa-dmi-wg

[26] Data Seal of Approval: http://www.datasealofapproval.org/

[27] Database Access and Integration Services WG (DAIS-WG)(OGF): http://forge.gridforum.org/projects/dais-wg/

[28] DEISA: http://www.deisa.eu/

[29] Digital Object Identifier (DOI): http://www.tib-hannover.de/en/the-tib/news/news/id/114/

[30] Disk Pool Manager (DPM): https://twiki.cern.ch/twiki/bin/view/LCG/DataManagementTop/

[31] DRIVER e-Publications: http://www.driver-repository.eu/Enhanced-Publications.html

[32] Drupal CMS: http://drupal.org/

[33] DSpace: http://www.dspace.org/

[34] Dublin Core Metadata Initiative: http://dublincore.org/

[35] e-Infrastructures Reflection Group (e-IRG): http://www.e-irg.eu/

[36] Electronic Data Capture Systems (EDC): Welker, J.A., *Implementation of electronic data capture systems - barriers and solution*, Comp Clin Trials 2007; 28 (3), 329-336

[37] Electronic Health Record (EHR): available from HIMSS website: http://www.himss.org/ASP/topics_ehr.asp [cited 2009 May 12]

[38] Electronic Health Records/Clinical Research (EHR/CR) project: eClinical Forum: EHR/CR (Electronic Health Records for Clinical Research) Project, April 1, 2008, available at: http://www.ehrcr.org/Docs/EHR_CR%20Funding%20Letter.pdf

[39] Elmroth, E., Hernandez, F., Tordsson, J., *Three Fundamental Dimensions of Scientific Workflow Interoperability: Model of Computation, Language, and Execution Environment*, Technical Report UMINF 09.05, Dept. of Computing Science and HPC2N, Umea University, submitted for publication, 2009, http://www.cs.umu.se/research/uminf/reports/2009/005/part1.pdf

[40] Enabling Grids for E-sciencE (EGEE): http://public.eu-egee.org/

[41] Enabling Grids for EsciencE: http://www.eu-egee.org/

[42] ePrints: http://www.eprints.org/

[43] eSciDoc: http://www.escidoc.org/

[44] ESTAT (2004): Results Task Force XIA, Extended Impact Assessment of INSPIRE, Eurostat, Hans Dufourmont (ed.): http://inspire.jrc.ec.europa.eu/reports/inspire_extended_impact_assessment.pdf

[45] Ethernet (IEEE 802): http://standards.ieee.org/getieee802/portfolio.html

[46] European Clinical Research Infrastructure Network (ECRIN): Demotes-Mainard, J., Ohmann, C., *European Clinical Research Infrastructures Network: promoting harmonisation and quality in European clinical research*, Lancet, 2005; Jan 8-14;365(9454): 107-108

[47] European Space Agency (ESA): http://www.esa.int/esaCP/index.html

[48] European Strategy Forum on Research Infrastructures (ESFRI): http://cordis.europa.eu/esfri/

[49] Extensible Markup Language (XML): http://www.w3.org/XML/

[50] Federated Identity Management: http://en.wikipedia.org/wiki/Federated_identity

[51] Fedora Commons: http://www.fedora-commons.org/

[52] FFmpeg: http://www.ffmpeg.org/

[53] Firewire (IEEE 1394-1995): http://standards.ieee.org/micro/1394overview.html

[54] Futurebus+ (IEEE 896-1991): http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00035051

[55] gLite File Transfer Service (FTS): https://twiki.cern.ch/twiki/bin/view/EGEE/FTS/

[56] Global EHR/CR Functional Profile Project: EHR/CR Functional Profile Working Group: Electronic Health Records/Clinical Research EHR/CR, Global Project and Profile Description Document, August 2007, available at: http://www.ehrcr.org/Docs/EHR-CR_Functional_Profile.pdf

[57] Global Spatial Data Infrastructure Association (GSDI): http://www.gsdi.org/

[58] Globus Reliable File Transfer (RFT): http://globus.org/toolkit/data/rft/

[59] Globus Replica Location Service (RLS): http://www.globus.org/rls/

[60] Globus: http://www.globus.org/

[61] Grid File Transfer Protocol (GridFTP)(OGF): http://www.ogf.org/documents/GFD.47.pdf

[62] Grid Security Infrastructure (GSI): http://www.globus.org/security/overview.html

[63] Ground European Network for Earth Science Interoperations - Digital Repositories (GENESI-DR): http://www.genesi-dr.eu/

[64] H.264: http://www.itu.int/rec/T-REC-H.264/e

[65] Handles System, Corporation for National Research Initiatives (CNRI): http://www.handle.net/

[66] Health Level 7 (HL7): available from Health Level Seven, Inc. homepage: http://www.hl7.org/ [cited 2009 May 12]

[67] HTTPS: http://www.w3.org/Protocols/rfc2616/rfc2616.html, and also http://www.openssl.org/

[68] iCAT project: http://code.google.com/p/icatproject/

[69] ICD-10-GM: http://www.who.int/classifications/apps/icd/icd10online/

[70] ICD9 diagnostic coding: http://en.wikipedia.org/wiki/List_of_ICD-9_codes

[71] Infrastructure for Spatial Information in Europe (INSPIRE): http://www.inspire-geoportal.eu/

[72] INSPIRE Directive (2007/2/EC): http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:0014:EN:PDF

[73] INSPIRE open letter: http://inspire.jrc.ec.europa.eu/openletter/

[74] Institution of Electrical and Electronics Engineers (IEEE): http://www.ieee.org/

[75] International Classification of Diseases (ICD): http://www.who.int/classifications/icd/en/index.html

[76] International Grid Trust Federation: http://www.igtf.net/

[77] International Standards Organization (ISO): http://www.iso.org/iso/home.htm

[78] International Virtual Observatory Alliance (IVOA): http://www.ivoa.net

[79] IPCC (2007): Climate Change 2007: Synthesis Report. Summary for Policymakers: http://www.ipcc.ch/pdf/assessment-report/ar4/syr/ar4_syr_spm.pdf

[80] iRODS: https://www.irods.org/

[81] ISLE Meta Data Initiative (IMDI): http://www.mpi.nl/IMDI/

[82] ISO-10746: Information technology - Open Distributed Processing - Reference model: Overview: http://standards.iso.org/ittf/PubliclyAvailableStandards/c020696_ISO_IEC_10746-1_1998(E).zip

[83] ISO-11179: http://metadata-standards.org/11179/

[84] ISO-12620: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=2517

[85] ISO-14721: Space data and information transfer systems – Open archival information system – Reference model: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683

[86] ISO-15836 (Dublin Core): http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=37629

[87] ISO-19101: Geographic information - Reference model: http://www.isotc211.org/Outreach/Overview/Factsheet_19101.pdf

[88] ISO-19107: Geographic information - Spatial schema: http://www.isotc211.org/Outreach/Overview/Factsheet_19107.pdf

[89] ISO-19108: Geographic information - Temporal schema: http://www.isotc211.org/Outreach/Overview/Factsheet_19108.pdf

[90] ISO-19109: Geographic information - Rules for application schema: http://www.isotc211.org/Outreach/Overview/Factsheet_19109.pdf

[91] ISO-19110: Geographic information - Methodology for feature cataloguing: http://www.isotc211.org/Outreach/Overview/Factsheet_19110.pdf

[92] ISO-19111: Geographic information - Spatial referencing by coordinates: http://www.isotc211.org/Outreach/Overview/Factsheet_19111.pdf

[93] ISO-19115: Geographic information - Metadata: http://www.isotc211.org/Outreach/Overview/Factsheet_19115.pdf

[94] ISO-19115: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=26020

[95] ISO-19123: Geographic information - Schema for coverage geometry and functions: http://www.isotc211.org/Outreach/Overview/Factsheet_19123.pdf

[96] ISO-19128: Geographic information - Web Map server interface: http://www.isotc211.org/Outreach/Overview/Factsheet_19128.pdf

[97] ISO-19139: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32557

[98] ISO-19142: Geographic information - Web Feature Service: http://www.isotc211.org/Outreach/Overview/Factsheet_19128.pdf

[99] ISO TC37/SC4: http://www.tc37sc4.org/

[100] ISO Technical Committee 211 (TC211): http://www.isotc211.org/

[101] ISO/IEC-14481: Information technology - Conceptual Schema Modelling Facilities (CSMF), see, for example: http://www.aim.nl/weblog/19981200%20ISO_FCD_14481.php

[102] ISOcat: http://www.isocat.org/

[103] IVOA Documents and Standards: http://www.ivoa.net/Documents/

[104] IVOA guidelines for participation: http://www.ivoa.net/Documents/latest/IVOAParticipation.html

[105] IVOA Ontology of astronomical object types: http://www.ivoa.net/Documents/latest/AstrObjectOntology.html

[106] IVOA Vocabularies: http://www.ivoa.net/Documents/latest/Vocabularies.html

[107] James, D., *Multiplexed Buses: The Endian Wars Continue*, IEEE Micro, Vol.10, No.3, pp.9-21, June 1990.

[108] Jensen, J., et al: *Practical Grid Storage Interoperation*, J. Grid Comp. special issue on interoperation, to appear.

[109] Kepler: http://www.kepler-project.org/

[110] Koehler, W., *A longitudinal study of Web pages continued: a consideration of document persistence*, J.Information Research, Vol.9, No.2, January 2004, http://informationr.net/ir/9-2/paper174.html

[111] Laboratory Data Model (LAB): available from CDISC website: http://www.cdisc.org/models/lab/v1.0.1/index.html

[112] LAMUS: Broeder, D., Claus, A., Offenga, F., Skiba, R., Trilsbeek, P., and Wittenburg, P., *LAMUS: The Language Archive Management and Upload System*, In Proc.5th Int.Conf.Language Resources and Evaluation (LREC 2006) (pp. 2291-2294), 2006; also see: http://www.clarin.eu/tools/lamus/

[113] Large Hadron Collider (LHC): http://public.web.cern.ch/public/en/LHC/LHC-en.html

[114] LCG File Catalog (LFC): https://twiki.cern.ch/twiki/bin/view/LCG/DataManagementDocumentation/

[115] Learning Object Metadata (LOM): http://de.wikipedia.org/wiki/Learning_Objects_Metadata

[116] Legal Issues in the e-Infrastructure, e-IRG Support Programme (e-IRGSP2), 15 January 2009, http://www.e-irg.eu/images/stories/publ/e-irgps2_public_deliverables/e-irgsp2_d4.1.pdf

[117] Lexical Markup Framework: http://www.lexicalmarkupframework.org/

[118] LEXUS: http://www.lat-mpi.eu/tools/lexus

[119] LHC Computing Grid (LCG): http://lcg.web.cern.ch/LCG/

[120] linear PCM: http://www.digitalpreservation.gov/formats/fdd/fdd000011.shtml

[121] LOINC: Huff, S.M., Rocha, R.A., McDonald C.J., et.al., *Development of the Logical Observation Identifier Names and Codes (LOINC) Vocabulary*, Journal of the American Medical Informatics Association. 1998, 5:276-292.

[122] MedDRA: vailable from homepage of Medical Dictionary for Regulatory Activities: http://www.meddramsso.com/MSSOWeb/index.htm [cited 2009 May 12]

[123] Metadata and Quality, e-IRG Data Management Task Force Report, June 2009.

[124] Metadata Encoding and Transmission Standard (METS): http://www.loc.gov/standards/mets/

[125] Moving Picture Experts Group (MPEG) standards (MPEG-1/2/3/4): http://www.chiariglione.org/mpeg/

[126] MPEG-1 Audio Layer 3 (MP3) (not to be confused with MPEG-3): http://en.wikipedia.org/wiki/MP3

[127] MPEG21 DID: http://www.chiariglione.org/MPEG/technologies/mp21-did/index.htm

[128] MPICH-G2: http://www3.niu.edu/mpi/

[129] MPICH1: http://www.mcs.anl.gov/research/projects/mpi/mpich1/

[130] MPICH2: http://www.mcs.anl.gov/research/projects/mpich2/

[131] Nature (2007): Climate change 2007: What They're Saying, Nature vol.445, p.579 (8 February 2007), doi:10.1038/445579a, http://www.nature.com/nature/journal/v445/n7128/pdf/445579a.pdf

[132] NHIN Slipstream Project: Walsh, L., Apathy, J., *NHIN Slipstream Project*, presentation at the AMIA Meeting on Pharmacovigilance, June 13, 2007

[133] OGF Grid Interoperability Now (GIN): http://www.ogf.org/gf/group_info/view.php?group=gin-cg

[134] OGF GridFTP: http://forge.gridforum.org/projects/gridftp-wg/

[135] OGF Repository Working Group: http://www.ogf.org/gf/group_info/view.php?group=dr-rg

[136] OGSA-DAI: http://www.ogsadai.org.uk/

[137] OGSA-DAI: http://www.ogsadai.org.uk/

[138] Ohmann, O., Kuchinke, W.: *Future developments of medical informatics from the viewpoint of networked clinical research. Interoperability and integration*, Methods Inform Med., Vol.48, pp.45-54, 2009.

[139] Online Computer Library Center (OCLC) Digital Repository Certification: http://www.oclc.org/programs/ourwork/past/repositorycert.htm

[140] Open Archival Information System (OAIS): http://public.ccsds.org/publications/archive/650x0b1.pdf

[141] Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH): http://www.openarchives.org/

[142] Open Archives Initiative Object Reuse and Exchange (OAI-ORE): http://www.openarchives.org/ore/

[143] Open Geospatial Consortium (OGC): http://www.opengeospatial.org

[144] Open Geospatial Consortium (OGC): http://www.opengeospatial.org/

[145] Open Grid Forum (OGF): http://www.ogf.org/

[146] Open Language Archives Community (OLAC): http://www.language-archives.org/

[147] Open Metadir 2 (OM2): http://www.openmetadir.org/files/om2-architecture-intro.pdf

[148] OpenEHR Reference Model (RM): available from the openEHR Java Reference Implementation Project: http://www.openehr.org/projects/java.html [cited 2009 May 12]

[149] OpenEHR: http://www.openehr.org/home.html [cited 2009 May 12]

[150] OpenID: http://openid.net/

[151] OpenMPI: http://www.open-mpi.org/

[152] OpenRepositories Community: http://www.openrepositories.org/

[153] OpenScienceGrid (OSG): http://www.opensciencegrid.org/

[154] OpenSearch: http://www.opensearch.org/Home

[155] Operational Data Model (ODM): available from CDISC website: http://www.cdisc.org/models/odm/v1.3/index.html[cited 2009 May 12]

[156] OPS: available from: German Procedure Classification OPS: http://www.dimdi.de/static/en/klassi/prozeduren/ops301/index.htm [cited 2009 May 12]

[157] Organization for the Advancement of Structured Information Standards (OASIS): http://www.oasis-open.org/

[158] Overview of the International Virtual Observatory Alliance (IVOA): http://www.ivoa.net/pub/info/TheIVOA.pdf

[159] Pathology Coding: Pathology/Lab Coder's Survival Guide, The Coding Institute, 2009

[160] POH, Oncology and Aids networks of excellence: available at the Competence Networks in Medicine website: http://www.kompetenznetze-medizin.de/eng/_html/_ie/_start_ie.htm [cited 2009 May 12]

[161] Portable Operating System Interface for Computer Environments (POSIX) (IEEE 1003.1-1988): http://ieeexplore.ieee.org/xpls/abs_all.jsp?tp=&isnumber=4708&arnumber=182902&punumber=2893

[162] PREservation Metadata: Implementation Strategies (PREMIS): http://www.oclc.org/research/projects/pmwg/

[163] Preserving Access to Digital Information (PADI): http://www.nla.gov.au/padi/topics/36.html

[164] Protocol Representation Group (PRG): available from CDISC website: http://www.cdisc.org/standards/protocol.html [cited 2009 May 12]

[165] RBRVS Values: The Essential RBRVS: A Comprehensive Listing of RBRVS Values for CPT and HCPCS Codes, publisher Ingenix, 2009

[166] Really Simple Syndication (RSS): http://www.rssboard.org/rss-specification

[167] RedIRIS: http://www.rediris.es/

[168] Reference Information Model (RIM): available from HL7 webpage: http://www.hl7.org/Library/data-model/RIM/modelpage_mem.htm [cited 2009 May 12]

[169] Representation State Transfer (REST): Roy Fielding, *Architectural Styles and the Design of Network-based Software Architectures*, PhD. Dissertation, 2000: http://www.ics.uci.edu/ fielding/pubs/dissertation/top.htm

[170] Resource Description Framework (RDF/XML): http://www.w3.org/TR/rdf-syntax-grammar/

[171] Scalable Coherent Interconnect (SCI) (IEEE 1596-1992): http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00347683

[172] Simple Object Access Protocol (SOAP): http://www.w3.org/TR/soap/

[173] SNOMED CT: available from the International Health Terminology Standards Development Organisation website: http://www.ihtsdo.org/snomed-ct/ [cited 2009 May 12]

[174] Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., and Miller, H.L. (eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

[175] Standards for Exchange of Non-clinical Data (SEND): available from CDISC website: http://www.cdisc.org/models/send/v2.3/index.html [cited 2009 May 12]

[176] Storage Networking Industry Association (SNIA): http://www.snia.org/

[177] Storage Resource Broker (SRB): http://www.sdsc.edu/srb/index.php/Main_Page

[178] Storage Resource Manager (SRM protocol/interface): http://www.gridpp.ac.uk/wiki/SRM/

[179] Storage Resource Manager (StoRM): http://storm.forge.cnaf.infn.it/

[180] Study Data Tabulation Model (SDTM): available from CDISC website: http://www.cdisc.org/models/sdtm/v1.1/index.html [cited 2009 May 12]

[181] Survey of Health, Ageing and Retirement in Europe (SHARE): http://www.share-project.org/

[182] Taverna: http://taverna.sourceforge.net/

[183] Telematics Platform (TMF): http://www.gesundheitsforschung-bmbf.de/en/583.php

[184] Text Encoding Initiative (TEI): http://www.tei-c.org/

[185] TextGrid project: http://www.textgrid.de/

[186] The AstroGrid (UK-VO Technology Centre) project: http://www.astrogrid.org/

[187] The European Virtual Observatory: http://www.euro-vo.org/pub/

[188] The International Astronomical Union (IAU) Division XII Commission 5 on Documentation & Astronomical Data: http://www.iau.org/science/scientific_bodies/commissions/5/

[189] The IVOA in 2008: Assessment and Future Roadmap: http://www.ivoa.net/Documents/latest/IVOARoadMap-2008.html

[190] The IVOA Newsletter: http://www.ivoa.net/newsletter/

[191] TNM: Sobin, L.H., Greene, F.L.: *TNM classification*, Canver 2001; 92 (2): 452

[192] Transactional Deployment System (TDS): http://www.springerlink.com/content/3ey8vmgjx1ehgft0/

[193] UNICODE: http://unicode.org/

[194] Unified Modeling Language (UML): http://www.uml.org/

[195] Uniform Resource Identifier (URI): http://www.w3.org/Addressing/

[196] Uniform Resource Name (URN): http://www.w3.org/TR/uri-clarification/

[197] United Nations Framework Convention on Climate Change (UNFCCC): http://unfccc.int/

[198] Universal Content Descriptors (UCDs): http://www.ivoa.net/Documents/latest/UCD.html

[199] Universal Description, Discovery, and Integration (UDDI): http://uddi.xml.org/

[200] Universal Serial Bus (USB): http://www.usb.org/

[201] VOEvent transient event protocol: http://www.ivoa.net/cgi-bin/twiki/bin/view/IVOA/IvoaVOEvent

[202] VP-Core: http://www.vpcore.nl/

[203] WebCom-G: http://www.webcom-g.org/

[204] Workflows in astronomy by python scripting:
http://www.astrogrid.org/wiki/Help/IntroScripting/AstrogridPython

[205] World Wide Web Consortium (W3C): http://www.w3c.org/

[206] WS-Interoperability Organisation: http://www.ws-i.org/

[207] WSDL: http://www.w3.org/TR/wsdl/

Appendix A

Data Seal of Approval (DSA) Overview

1 The Data Seal of Approval

The Data Seal of Approval was established by a number of institutions committed to the long-term archiving of research data. By assigning the seal, the DSA group seeks to guarantee the durability of the data concerned, but also to promote the goal of durable archiving in general.

The Data Seal of Approval is granted to repositories that are committed to archiving and providing access to scholarly research data in a sustainable way. It is assigned by the DSA Assessment Editorial Board and renewed every year through a modification procedure.

The Editorial Board consists of members from the following institutes: Alfred Wegener Institute (Germany), CINES (France), DANS (The Netherlands), ICPSR (USA), MPI for Psycholinguistics (The Netherlands), NESTOR (Germany) and UK Data Archive (United Kingdom).

Institutions that have achieved the Data Seal of Approval have the right to display the DSA logo on their Web sites, thereby demonstrating the reliability of their archival processes and procedures, without this entailing new thresholds, regulations or high costs.

2 The DSA Assessment

Achieving the DSA means that the data archive or repository is in compliance with the sixteen DSA guidelines, as determined through an assessment procedure. Although these guidelines pertain to three stakeholders – the data producer (three guidelines), the data consumer (three guidelines) and the data archive (ten guidelines) – the data archive is seen as the primary implementer of the guidelines. The data archive as an organization should assume responsibility for the overall implementation of the DSA in its own specific field.

3 Procedure

To achieve the Data Seal of Approval, and thereby receive permission to use the DSA logo, the repository must keep a file directory on the Web that is accessible through the front page of the repository. This assessment directory contains:

1. An up-to-date version of the Data Seal of Approval handbook (datasealofapproval.pdf)
2. The information leaflet about the Data Seal of Approval Assessment (aboutDSAA.pdf)
3. The Data Seal of Approval Assessment form (DSAA.pdf)

All the above-mentioned documents can be found on the DSA Web site: http://www.datasealofapproval.org

The completion of the DSA self-assessment form is the starting point for the reviewing procedure. The assessment is then sent to the Editorial Board in order to decide via peer review whether an organization will be granted the Seal of Approval. There is no audit, no certification: just a review on the basis of trust.

The DSAA.pdf consists of a guide to facilitate completion of the assessment, the assessment itself and a modification record.

- The assessment lists the sixteen Data Seal of Approval guidelines. In the assessment, the organization describes how these guidelines relate to the repository and how they have been implemented. The assessment reflects the current situation of the repository in a transparent and open manner.
- In the modification record, every modification made to the assessment is recorded. If information changes, the organization informs the DSAA EB by email info@datasealofapproval.org

When approval is granted by the DSA Assessment Editorial Board, the DSA logo is displayed on the front page of the repository by means of HTML code, which the organization receives from the Board. It contains a link to http://www.datasealofapproval.org as well as a link to the organization's repository assessment directory.

The Editorial Board places a link pointing to the repository on the Web site http://datasealofapproval.org, using the logo of the specific repository or the name of the repository in combination with the logo of the hosting organization.

4 THE DATA SEAL OF APPROVAL GUIDELINES

1. The *data producer* deposits the research data in a data repository with sufficient information for others to assess the scientific and scholarly quality of the research data and compliance with disciplinary norms.
2. The *data producer* provides the research data in formats recommended by the data repository.
3. The *data producer* provides the research data together with the metadata requested by the data repository.
4. The *data repository* has an explicit mission in the area of digital archiving and promulgates it.
5. The *data repository* uses due diligence to ensure compliance with legal regulations and contracts.
6. The *data repository* applies documented processes and procedures for managing data storage.
7. The *data repository* has a plan for long-term preservation of its digital assets.
8. Archiving takes place according to explicit workflows across the data life cycle.
9. The *data repository* assumes responsibility from the data producers for access to and availability of the digital objects.
10. The *data repository* enables the users to utilize the research data and refer to them.
11. The *data repository* ensures the integrity of the digital objects and the metadata.
12. The *data repository* ensures the authenticity of the digital objects and the metadata.
13. The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.
14. The *data consumer* must comply with access regulations set by the data repository.
15. The *data consumer* conforms to and agrees with any codes of conduct that are generally accepted in higher education and research for the exchange and proper use of knowledge and information.
16. The *data consumer* respects the applicable licences of the data repository regarding the use of the research data.