#1

COMPLETE

Collector: Web Link (Web Link)
Started: Monday, February 03, 2014 3:57:46 AM
Last Modified: Friday, February 07, 2014 6:44:58 AM
Time Spent: Over a day
IP Address: 193.11.31.253

PAGE 1

Q1: **Contact Information**

Name:	Ingemar Häggström
Organisation:	EISCAT
City/Town:	Kiruna
Country:	Sweden
Email Address:	ingemar@eiscat.se

Q2: **What is the research group or project for which you are providing information about research data lifecycle? Please, provide some background information on the size and geographical distribution of the research collaboration/project and about the main scientific challenges being addressed.**

EISCAT, the European Incoherent Scatter Scientific Association, is established to conduct research on the lower, middle and upper atmosphere and ionosphere using the incoherent scatter radar technique. This technique is the most powerful ground-based tool for these research applications. EISCAT is also being used as a coherent scatter radar for studying instabilities in the ionosphere, as well as for investigating the structure and dynamics of the middle atmosphere and as a diagnostic instrument in ionospheric modification experiments with the Heating facility.
The EISCAT Scientific Association is funded and operated by research councils of Norway, Sweden, Finland, Japan, China and the United Kingdom (collectively, the EISCAT Associates).

Q3: **What are the scientific disciplines covered by your research activities? (select one or more from the EGI scientific discipline classification)**

Discipline 1	1.6.25 Space science
Discipline 2	1.6.24 Plasma physics
Discipline 3	1.4.1 Atmospheric science

Q4: **Data products: what are the data sources, what type of data formats and what is the amount of data produced including the frequency of generation/update?**

Incoherent Scatter Radar data in a different refinement levels
Packed into HDF5
Eternal storage 2PB/y

Q5: **Data Storage: what digital data is retained during the life of the project/experiment? In what media and location?**

Selected low level data
Completness increases for higher levels

Q6: **Data Management Plans: Do you have a well defined data management plan?** No

Q7: **Meta data collection, management, and packaging: Please indicate, how you collect, manage, and package your meta data in order to prepare it for supporting archival and preservation of your research data and what type of other meta data you (do/would like to) store along with your research data. What metadata standards do you currently use? Please describe the technical challenges to be addressed.**

Metadata are from several sources. We are in process of packing the metada into complete hdf5 files for export with the data.

Q8: **Persistent Identifier: Please indicate, how you generate your persistent identifiers for your data?**

Currently we do not have PIDs for all (meta)data sources

Q9: **Preserved data: Please indicate the type(s) of data and the amount of data (GB per year, overall)**

HDF5, 2PB/y, 10PB

Q10: Support for access and re-use: How do you presently store your digital research data for future access and use? What kinds of future re-use cases could you envisage? How much of your data do you make available to others? (none, some, most, all) If all or part of your data are not available to others, why not? What type of access policy is in place (e.g., open access)? What type of security is implemented? (e.g., open, local password, LDAP, X.509 certificate)

Currently on online disks, to be changed to tape storage.
High factor of reuse as the data is continous
Limited access for lower levels of data

Q11: Business models: What is the current business model for the gathering, archival, preservation, and re-use of the data?

NA

Q12: Can you describe if/what educational and training needs you could identify in any of the areas mentioned above? This will very important for us so to define future educational and training programs.

High threashold to be able to define searches, so education is much needed

Q13: Would you in general be willing to present a use case/scenario/requirement/demand from the above in the workshop

Yes

#2

COMPLETE

Collector: Web Link (Web Link)
Started: Thursday, February 20, 2014 12:03:57 AM
Last Modified: Thursday, February 20, 2014 12:49:02 AM
Time Spent: 00:45:04
IP Address: 2.86.158.216

PAGE 1

Q1: **Contact Information**

Name:	Andreas Drakos
Organisation:	University of Alcala
City/Town:	Alcala de Henares
Country:	Spain
Email Address:	drakos@agroknow.gr

Q2: **What is the research group or project for which you are providing information about research data lifecycle? Please, provide some background information on the size and geographical distribution of the research collaboration/project and about the main scientific challenges being addressed.**

The related project is the agINFRA FP7 project (www.aginfra.eu). It is a project aggregating agricultural research data available worldwide. It aims to provide the e-infrastructure which will facilitate the exchange of knowledge of the research agricultural community. At the same time the project provides tools and services for aggregating a large number of metadata related to the agricultural community as well as interlinking them through a linked open agricultural data layer proposed by the project.

Q3: **What are the scientific disciplines covered by your research activities? (select one or more from the EGI scientific discipline classification)**

Discipline 1	1.3.4 Information management
Discipline 2	4.1.1 Agriculture

Q4: **Data products: what are the data sources, what type of data formats and what is the amount of data produced including the frequency of generation/update?**

Data sources: Institutional repositories, metadata aggregators,
Data formats: PDF, HTML, multimedia (videos, images), XLS/CSV, maps
Data types: Publications / Bibliographic references, educational resources, germplasm data, soil maps, statistical data, researchers' profiles
Amount of data harvested/aggregated: 10M bibliographic/educational metadata records

Q5: **Data Storage: what digital data is retained during the life of the project/experiment? In what media and location?**

The project provides cloud-based tools and services for institutions and individuals to create and manage digital data repositories, focusing more on bibliographic and educational repositories.

Q6: **Data Management Plans: Do you have a well defined data management plan?** No

Q7: **Meta data collection, management, and packaging: Please indicate, how you collect, manage, and package your meta data in order to prepare it for supporting archival and preservation of your research data and what type of other meta data you (do/would like to) store along with your research data. What metadata standards do you currently use? Please describe the technical challenges to be addressed.**

The project has provided metadata management workflows for the aggregation of metadata through harvesting (OAI-PMH) and ingestion (XML dumps). Different workflows have been used for different data types. Metadata are stored in the agINFRA cloud and processed using the agINFRA grid services.

The metadata standards used include AGRIS AP, Dublin Core, METS, MODS, IEEE LOM, Organic.Edunet IEEE LOM AP, INSPIRE-based metadata schema (soil maps), EURISCO Descriptors (Germplasm), SDMX (statistical), LOD (learning/bibliographic)

The wide variety of the metadata schemas used in the digital repositories to be aggregated by agINFRA led to the need for developing an abstract metadata schema, meeting the individual requirements of the existing schemas. In some cases, the repositories are not fully compliant with a standard metadata schema, requiring specific transformations. In other cases, the quality of metadata is poor (e.g. low completeness) leading to the need for semi-automatic metadata enrichment.

Q8: **Persistent Identifier: Please indicate, how you generate your persistent identifiers for your data?**

Identifiers are created automatically but they do not follow any specific standard.

Q9: **Preserved data: Please indicate the type(s) of data and the amount of data (GB per year, overall)**

Since the project provides cloud-based tools for the management of digital repositories, it is not possible to provide an estimation for this question.

Q10: **Support for access and re-use: How do you presently store your digital research data for future access and use? What kinds of future re-use cases could you envisage? How much of your data do you make available to others? (none, some, most, all) If all or part of your data are not available to others, why not? What type of access policy is in place (e.g., open access)? What type of security is implemented? (e.g., open, local password, LDAP, X.509 certificate)**

In several cases, digital research data are stored in digital repositories along with their metadata. Automatic processes can be enabled for the exposure of these data as linked open data or to be harvested using the agINFRA metadata management workflows.

The agINFRA requires its data providers to be compliant with the open data movement, therefore data provided through agINFRA are open access.

Q11: **Business models: What is the current business model for the gathering, archival, preservation, and re-use of the data?**

The agINFRA project aims to provide all the tools and services for the agricultural community to be able to create and manage digital repositories. The project currently is in the process of creating a business model and sustainability plans in order to support its main vision.

Q12: **Can you describe if/what educational and training needs you could identify in any of the areas mentioned above? This will very important for us so to define future educational and training programs.**

In the agricultural community there is a huge gap between agricultural researchers and the technology used regarding digital repositories and uses of data, The main problem is to engage this community and train them in using the tools and services that are available for serving their purposes.

Q13: **Would you in general be willing to present a use case/scenario/requirement/demand from the above in the workshop**

Yes, during the agINFRA project a number of use cases have been developed and may be presented during the workshop if needed.

#3

COMPLETE

Collector: Web Link (Web Link)
Started: Thursday, February 20, 2014 4:42:55 AM
Last Modified: Thursday, February 20, 2014 7:12:25 AM
Time Spent: 02:29:30
IP Address: 130.246.132.177

PAGE 1

Q1: **Contact Information**

Name:	Matthew Viljoen
Organisation:	STFC RAL
City/Town:	Didcot
Country:	UK
Email Address:	matthew.viljoen@stfc.ac.uk

Q2: **What is the research group or project for which you are providing information about research data lifecycle? Please, provide some background information on the size and geographical distribution of the research collaboration/project and about the main scientific challenges being addressed.**

We provide a Digital Preservation Service for the ISIS Pulsed Neutron and Muon source. ISIS is situated at the Rutherford Appleton Laboratory in the UK and supports the national and international community of more than 3000 scientists for research into subjects ranging from clean energy and the environment, pharmaceuticals and health care, through to nanotechnology and materials engineering, catalysis and polymers, and on to fundamental studies of materials.

Q3: **What are the scientific disciplines covered by your research activities? (select one or more from the EGI scientific discipline classification)**

Discipline 1	1.4 Earth sciences
Discipline 2	1.5 Biological sciences
Discipline 3	1.7 Chemical sciences

Q4: **Data products: what are the data sources, what type of data formats and what is the amount of data produced including the frequency of generation/update?**

Examples are: Muon spectroscopy, Neutron diffraction and spectroscopy, Reflectometry and Small angle scattering

Q5: **Data Storage: what digital data is retained during the life of the project/experiment? In what media and location?**

All data created by ISIS is preserved using Safety Deposit Box (SDB) by Tessela on an ORACLE StorageTek SL8500 tape robot using DMF by SGI on T10KA media.

Q6: **Data Management Plans: Do you have a well defined data management plan?**

Yes,

If yes, add the URL
http://www.isis.stfc.ac.uk/user-office/data-policy11204.html

Q7: **Meta data collection, management, and packaging: Please indicate, how you collect, manage, and package your meta data in order to prepare it for supporting archival and preservation of your research data and what type of other meta data you (do/would like to) store along with your research data. What metadata standards do you currently use? Please describe the technical challenges to be addressed.**

After collection, data is initially stored on a Windows filestore. From there data is ingested into SDB. Metadata is created and stored in an ICAT instance (http://code.google.com/p/icatproject/) using the Core Scientific Metadata Model (CSMD)

Q8: **Persistent Identifier: Please indicate, how you generate your persistent identifiers for your data?**

DataCite DOIs may be generated on demand using web applications and a landing page developed around ICAT.

Q9: **Preserved data: Please indicate the type(s) of data and the amount of data (GB per year, overall)**

Data is stored in the NeXus format (common data format for neutron, x-ray and muon sciences). To date we have approximately 11TB of data stored.

Q10: Support for access and re-use: How do you presently store your digital research data for future access and use? What kinds of future re-use cases could you envisage? How much of your data do you make available to others? (none, some, most, all) If all or part of your data are not available to others, why not? What type of access policy is in place (e.g., open access)? What type of security is implemented? (e.g., open, local password, LDAP, X.509 certificate)

The archive described above is for long term digital preservation - and to ease the migration of data to evolving file formats. Data for access is additionally stored on the Windows filestore.

Regarding open access, as specified in the ISIS data management policy:
3.1.1 All raw data and the associated metadata obtained as a result of free (non-commercial) access to ISIS, reside in the public domain, with ISIS acting as the custodian.
3.1.2 All raw data and the associated metadata obtained as a result of 'commercial-in-confidence' access to ISIS will be owned exclusively by the commercial user. Commercial users must agree with their relevant instruments scientists how they wish their raw data and metadata to be managed before the start of any experiment.

Also:
3.3.1 Access to raw data and metadata beyond the period that it is stored on instrument-related computers will be via a searchable on-line catalogue.
3.3.2 Access to the on-line catalogue will be restricted to those who register with STFC/ISIS as users of the on-line catalogue.

Q11: Business models: What is the current business model for the gathering, archival, preservation, and re-use of the data?

Respondent skipped this question

Q12: Can you describe if/what educational and training needs you could identify in any of the areas mentioned above? This will very important for us so to define future educational and training programs.

Unresolved issues we face include the best way to do hierarchical referencing of Digital Object Identifiers. i.e. relations between different versions of data if the data changes

Q13: Would you in general be willing to present a use case/scenario/requirement/demand from the above in the workshop

Yes

#4

COMPLETE

Collector: Web Link (Web Link)
Started: Thursday, February 20, 2014 6:35:06 AM
Last Modified: Thursday, February 20, 2014 8:53:52 AM
Time Spent: 02:18:46
IP Address: 217.26.87.7

PAGE 1

Q1: Contact Information

Name:	Luigi Carotenuto
Organisation:	Telespazio s.p.a.
City/Town:	Naples
Country:	Italy
Email Address:	luigi.carotenuto@telespazio.com

Q2: What is the research group or project for which you are providing information about research data lifecycle? Please, provide some background information on the size and geographical distribution of the research collaboration/project and about the main scientific challenges being addressed.

I coordinated the ULISSE and CIRCE projects, driven by Telespazio, CNES and DLR and supported by the network of the 8 European centres for scientific operations on ISS (USOC).
Currently science data produced on ISS is distributed to scientists responsible for the experiments, without a systematic exploitation in the long term. Consequently, ISS datasets are distributed, heterogeneous and complex, even is the total volume is not large.
Distribution of ISS data to scientific community is subjected to constraints and requires authorization.
ULISSE developed a demonstrator of a service platform for ISS data preservation and controlled dissemination.
CIRCE as coordination action produced a roadmap for the establishment of a platform for ISS data preservation.
Main challenges concern:
o involving large user communities, not aware of space activities and belonging to different scientific domains
o providing comprehensive metadata to enable actual data re-analysis
o promoting cross-fertilization among domains
o ensuring the protection of intellectual property and privacy (for sensitive clinical data) through a rigorous management of user authentication, authorization and accounting (AAA) processes
CIRCE consortium intends to produce exploitable datasets and metadata, enhancing space data visibility toward the users and supporting its preservation and exploitation. To this aim the CIRCE consortium would like to explore the opportunity of a possible cooperation with existing infrastructures as EGI for the provision of services for data preservation (including ID provision) and user AAA.

Q3: What are the scientific disciplines covered by your research activities? (select one or more from the EGI scientific discipline classification)

Discipline 1	3 Medical and Health Sciences
Discipline 2	1.5 Biological sciences
Discipline 3	1.6 Physical sciences

Q4: Data products: what are the data sources, what type of data formats and what is the amount of data produced including the frequency of generation/update?

Raw data consists of the telemetry received from ISS, formatted according to the CCSDS standard.
CIRCE consortium intends to process the telemetry to produce exploitable datasets in a standard format. Data format must be open and widely distributed.
An amount of the order of 50 TB is expected to be produced in about 10 years.
Exploitable data is generated offline with respect to space operations; data is subjected to embargo (at least 1 year) in which only the scientific team responsible for the experiment can have access to data.
In average, about 40 exploitable datasets should be created every year; metadata could be updated several times per year as needed (for instance to insert new links to newly available resources).

Q5: Data Storage: what digital data is retained during the life of the project/experiment? In what media and location?

All data and metadata are retained in the long term (at least 10 years).
There are no major specifications about media and location.

Q6: Data Management Plans: Do you have a well defined data management plan? No

Q7: Meta data collection, management, and packaging: Please indicate, how you collect, manage, and package your meta data in order to prepare it for supporting archival and preservation of your research data and what type of other meta data you (do/would like to) store along with your research data. What metadata standards do you currently use? Please describe the technical challenges to be addressed.

Metadata concerns science, space equipment and operations of the space experiments. The source of information is a set of documents (scientific requirements, drawing and design of space equipment, logs of operations) that are available at the space operation centres.
Metadata is compiled according to a schema (a draft is available) to generate XML files. Text mining tools may support metadata generation.
Metadata contains also links to datasets; in ULISSE these links bring the user to a form for data request (data request needs prior user registration).
In order to support cross-fertilization among domains, metadata must include additional information about the scientific domain, linking experiment features to more general scientific concepts of different domains; semantic approach is needed.
Main challenges are:
o diversity of domains makes difficult to identify a common schema
o a core schema (with basic metadata related to the dataset) is advisable to ensure interoperability
o further metadata could be compliant to domain-specific ontology
o it is necessary to select an adequate semantic technology, supporting an easy interconnection, navigation and enquiring

Q8: Persistent Identifier: Please indicate, how you generate your persistent identifiers for your data?

Persistent ID are needed, their standard is not defined yet.

Q9: Preserved data: Please indicate the type(s) of data and the amount of data (GB per year, overall)

All types of data are produced: text, images, video, tables, time series, organic and inorganic samples.
In average about 5.000 GB of data is generated per year (about 50.000 GB overall in the next 10 years).

Q10: Support for access and re-use: How do you presently store your digital research data for future access and use? What kinds of future re-use cases could you envisage? How much of your data do you make available to others? (none, some, most, all) If all or part of your data are not available to others, why not? What type of access policy is in place (e.g., open access)? What type of security is implemented? (e.g., open, local password, LDAP, X.509 certificate)

Datasets are stored as file systems. The file system is organized in folders, each one related to an experiment.
It is expected that data re-use will concern mainly scientific purposes and particularly for:
o future refinement of analysis with new tools/knowledge
o comparison with future complementary data
o merging with future homogeneous data to increase statistics
o re-analysis for different research objectives
All metadata is freely accessible without constraints. All datasets can be made available upon authorization. The reference data policy is the ESA Data Policy for Human Spaceflights.
The reference security is LDAP.

Q11: Business models: What is the current business model for the gathering, archival, preservation, and re-use of the data?

Preservation and exploitation of space data is presently a service of public utility: Space Agencies have the mission to maximize the return of space missions; space data users are mainly members of the scientific community (commercial use of space data is advisable but not mature yet). Possible financial support may be provided by Space Agencies (for basic services) and users (through the research grants on a pay-for-use basis).

Q12: Can you describe if/what educational and training needs you could identify in any of the areas mentioned above? This will very important for us so to define future educational and training programs.

Training may be advisable for:
o metadata compilation
o semantic links

Q13: Would you in general be willing to present a use case/scenario/requirement/demand from the above in the workshop

Yes

#5

COMPLETE

Collector: Web Link (Web Link)
Started: Monday, February 24, 2014 1:58:49 AM
Last Modified: Monday, February 24, 2014 3:03:08 AM
Time Spent: 01:04:18
IP Address: 145.23.254.101

PAGE 1

Q1: **Contact Information**

Name:	Alessandro Spinuso
Organisation:	KNMI
City/Town:	De Bilt
Country:	The Netherlands
Email Address:	spinuso@knmi.nl

Q2: **What is the research group or project for which you are providing information about research data lifecycle? Please, provide some background information on the size and geographical distribution of the research collaboration/project and about the main scientific challenges being addressed.**

Seismology addresses fundamental problems in understanding earthquake dynamics, seismic wave propagation and the properties of the Earth's subsurface at a large number of scales. These aim at aiding society in the forecasting and management of natural hazards, energy resources, environmental changes, and national security.
The VERCE project is supporting this effort by developing a data-intensive e-science environment to enable innovative data analysis and data modelling methods that fully exploit the increasing wealth of open data generated by the observational and monitoring systems of the global seismological community.
VERCE's strategy is to build upon a service-oriented architecture and a data-intensive platform delivering services, workflow tools, and software as a service, and to integrate the distributed European public data and computing infrastructures (GRID, HPC and CLOUD) with private resources and the European integrated data archives of the seismology community.

Q3: **What are the scientific disciplines covered by your research activities? (select one or more from the EGI scientific discipline classification)**

Discipline 1	1.4.5 Geophysics
Discipline 2	1.4.11 Seismology

Q4: **Data products: what are the data sources, what type of data formats and what is the amount of data produced including the frequency of generation/update?**

Let's consider as an example the requirements for the Forward Modelling use case, whose objective is to offer the possibility to perform simulations of seismic waves as a service to the seismological community. This use case is currently running within one of the EGI clusters (LRZ) and is controlled by a Science Gateway developed with the gUSE technology.
The data sources are configuration files and models which consist of roughly 300MB. Data types consists of binary (application specific), ascii (application specific), graphic visualization of output (png/pdf)
The amount of data produced depends from the setup of the experiment. In order to get started with the simulation, a typical scenario would produce initially a minimum amount of ~ 4GB of data. Eventually the simulation can produce thousands of synthetic seismograms of 240k each. Anyway, the highest frequency of production of data and medatada is concentrated in the post-processing phase, which can be indicated as the number-of-post-processing-steps/time-of-the-post-processing-computation in sec. That means, for instance, that if 100 stations will produce 900 products and metadata in 30 minutes, we might have a peak of one data product each two seconds.
The obtained synthetics will be used then for the evaluation of the model, by comparing them with real observations. This will require to access and download from the institutional archives at least the same amount of sensors' data.

Besides simulation, the project aims at providing also facilities for Data Intensive Computation in the field of Cross Correlation analysis. The current application is designed for an ingestion of 10 to 100 GB of input data. Intermediate data products are expected to be as large as the input dataset and those can be reused to evaluate different pre-processing techniques. Final results though are relatively small in size, we can consider them as 1/10 of the original Data. For what concerns the frequency, the description of the Cross Correlation use case aims to have full runs performed hourly, for time dependent variation analysis over near-real time observations

Q5: Data Storage: what digital data is retained during the life of the project/experiment? In what media and location?

All of the data produced from an experiment and the relative metadata should be preserved, unless it is indicated differently by the user. This will allow the scientists to try and evaluate different analysis techniques.
For the preservation of the data, our current system makes use of a federation of data archives (iRODS nodes). The federation will store observational datasets and data products obtained by the workflows' executions. The transfer of the results from the computing cluster to the data archives is performed only at the end of the computation (unfortunately local policies prevented us to do it differently) via a dedicated task, which establishes a gridftp connection between the cluster and the federation. Moreover, the cluster needs to be promptly cleaned after the processing because of the limited amount of space available. A consistent and effective clean-up policy is required which neither prevent other users from accessing the resource nor cause the loss of important results which, for instance, may be in the process of being transferred.
Moreover, a more dynamic and permissive approach to data transfer and visualisation, for instance, of provenance metadata, could indeed facilitate the rapid exploration of the results, which are often obtained after the execution of long and intensive runs.

Q6: Data Management Plans: Do you have a well defined data management plan? No

Q7: Meta data collection, management, and packaging: Please indicate, how you collect, manage, and package your meta data in order to prepare it for supporting archival and preservation of your research data and what type of other meta data you (do/would like to) store along with your research data. What metadata standards do you currently use? Please describe the technical challenges to be addressed.

The execution of the workflow's (jobs' executables in this case) pre and post processing steps produces provenance metadata which should be, in an ideal scenario, already available at runtime rather than at the end of the computation. These metadata provide immediate feedback to the users on the status of the processing, fostering pre- validation and, if needed, the interruption of the computation, saving CPU time and useless waits.
The current approach implemented within the EGI cluster hosted by LRZ, updates the provenance information to an external webservice at run time. The metadata vocabulary presents community terms and user annotations, with a particular attention towards the coverage of the W3C-PROV concepts for provenance representation. To the best of our knowledge, the runtime approach just described is not supported by other initiative like PRACE, which are preventing data exchange with external resources at any direction.
Other metadata are either collected after the transfer of the data from the cluster to the data federation nodes or at the time of the acquisition, when data are staged onto the data federation. These metadata are community defined and are extracted via the execution of microservices, which update a global metadata catalog hosted by a NoSql database. Our technological choice is motivated by the nature of the metadata schema, which is required to be dynamic and capable of supporting structured metadata and annotations.

Q8: Persistent Identifier: Please indicate, how you generate your persistent identifiers for your data?

The persistent identifiers are generated either during the execution of the workflow, or at the time of the acquisition of the data from the institutional archives.
Their generation relies on the working nodes of the hosting architecture. The identifiers get stored as part of the metadata describing the data products and their provenance, offering the possibility to link to the actual files stored within the data federation.

Q9: Preserved data: Please indicate the type(s) of data and the amount of data (GB per year, overall)

Being the VERCE a platform offering computational seismology as a service, the amount of preserved data will depend from the number of users and the number of relevant computations performed. Some of the Data Intensive use cases envisaged by the project (Cross correlation Analysis), might require to ingest observational datasets of the order of 10 TB, generating as much as intermediate data that needs to be preserved. This consideration assumes that users want to preserve only the intermediate products obtained by those runs that brought to relevant results.

Q10: Support for access and re-use: How do you presently store your digital research data for future access and use? What kinds of future re-use cases could you envisage? How much of your data do you make available to others? (none, some, most, all) If all or part of your data are not available to others, why not? What type of access policy is in place (e.g., open access)? What type of security is implemented? (e.g., open, local password, LDAP, X.509 certificate)

The access and re-use of the data produced will be in control of the user who runs the experiment, assuming some sort of intellectual property on the results.
Said that, quality preprocessed data could be reused by others in order to save computational time. (For instance, Cross correlation pre-processed data could be reused to evaluate different Cross correlation algorithms, while synthetics can be reused to apply different strategies to improve the initial model).
For these purposes metadata and provenance are important because they could provide a mean to express quality measurements on the datasets that are publicly accessible.
The transfer and the acess to our iRODS data federation is currently protected by password and x.509 certificate

Q11: Business models: What is the current business model for the gathering, archival, preservation, and re-use of the data?

NA

Q12: Can you describe if/what educational and training needs you could identify in any of the areas mentioned above? This will very important for us so to define future educational and training programs.

Training on data staging techniques and metadata/provenance management would be extremely useful once a global policy is defined. Moreover, it would be very important to inject these concepts into special trainings that should be designed for those engineers/computer-scientists who are trying to expose HPC/Data Intensive applications as a service for a specific community.

Q13: Would you in general be willing to present a use case/scenario/requirement/demand from the above in the workshop

YES

#6

COMPLETE

Collector: Web Link (Web Link)
Started: Monday, February 24, 2014 12:32:56 AM
Last Modified: Monday, February 24, 2014 5:00:03 AM
Time Spent: 04:27:06
IP Address: 130.246.132.178

PAGE 1

Q1: Contact Information

Name:	Matthew Viljoen
Organisation:	STFC RAL
City/Town:	Didcot
Country:	UK
Email Address:	matthew.viljoen@cern.ch

Q2: What is the research group or project for which you are providing information about research data lifecycle? Please, provide some background information on the size and geographical distribution of the research collaboration/project and about the main scientific challenges being addressed.

DPHEP is a study group focussing on data persistency and long term analysis for HEP and including LHC data from CERN. DPHEP aims to converge to a common set of specifications for this.

Q3: What are the scientific disciplines covered by your research activities? (select one or more from the EGI scientific discipline classification)

Discipline 1: 6.15 High energy physics

Q4: Data products: what are the data sources, what type of data formats and what is the amount of data produced including the frequency of generation/update?

LHC raw data produced at the detectors at CERN during data taking. Some 75PB of raw data so far at LHC in RAW file format. Analysis data in ROOT file formats. No updates; new data sets are produced (and the RAW data is immutable). Trigger rates range from 10Hz to a few GHz (across all HEP). Event sizes go from KB to GB.

Q5: Data Storage: what digital data is retained during the life of the project/experiment? In what media and location?

Data is archived on tape centrally at CERN and additionally at Tier 1s across the world using both disk and tape. Tape media is refreshed regularly onto new generations of media - typically every 3 years.

Q6: Data Management Plans: Do you have a well defined data management plan?

Yes,

If yes, add the URL
WLCG Technical Design Report - http://lcg-archive.web.cern.ch/lcg-archive/tdr.htm

Q7: Meta data collection, management, and packaging: Please indicate, how you collect, manage, and package your meta data in order to prepare it for supporting archival and preservation of your research data and what type of other meta data you (do/would like to) store along with your research data. What metadata standards do you currently use? Please describe the technical challenges to be addressed.

Data originates at LHC detectors where the selection of interesting event data in done. Primary Tier 0 archive of all data is at CERN. Further data replication is done by experiments across global network of Tier 1s according to each experiment's own data management policies.

Q8: Persistent Identifier: Please indicate, how you generate your persistent identifiers for your data?

No consistent way. Experiments are starting to adopt DOIs and there is increasing pressure to open data across HEP

Q9: Preserved data: Please indicate the type(s) of data and the amount of data (GB per year, overall)

RAW and ROOT file formats are the most pervasive. Currently approx. 25PB/year. This will be more in Run2 and more with every future run. It may be as much as 0.5EB/year in HL-LHC.

Q10: Support for access and re-use: How do you presently store your digital research data for future access and use? What kinds of future re-use cases could you envisage? How much of your data do you make available to others? (none, some, most, all) If all or part of your data are not available to others, why not? What type of access policy is in place (e.g., open access)? What type of security is implemented? (e.g., open, local password, LDAP, X.509 certificate)

Currently not all data is available publicly but this will change as experiments review their open data policies - which is becoming required by funding agencies and needs to be addressed soon. We will follow widely-used standards.

Q11: Business models: What is the current business model for the gathering, archival, preservation, and re-use of the data?

Tangible benefits by maximizing return of experiment investment to taxpayers by promoting research and verification of results through open access and reliable preservation of data.

Q12: Can you describe if/what educational and training needs you could identify in any of the areas mentioned above? This will very important for us so to define future educational and training programs.

Training should be broken down by activity. Training for bit preservation providers is very different to that of developers. The former can be done under existing bodies such as HEPiX etc.

Q13: Would you in general be willing to present a use case/scenario/requirement/demand from the above in the workshop

Yes

#7

COMPLETE

Collector: Web Link (Web Link)
Started: Monday, February 24, 2014 6:37:02 AM
Last Modified: Monday, February 24, 2014 7:18:56 AM
Time Spent: 00:41:54
IP Address: 129.175.127.172

PAGE 1

Q1: **Contact Information**

Name:	Karima Rafes
Organisation:	Inria Saclay
City/Town:	Orsay
Country:	France
Email Address:	karima.rafes@inria.fr

Q2: **What is the research group or project for which you are providing information about research data lifecycle? Please, provide some background information on the size and geographical distribution of the research collaboration/project and about the main scientific challenges being addressed.**

The information concerns the CENTER FOR DATA SCIENCE (CDS). CDS is a transversal interdisciplinary laboratory of the Paris Saclay University (UPSay) dedicated to data science, both fundamental and applied. UPSay is a superstructure for University Paris Sud, research institutions (iNRIA-Saclay, INRA, ...) and the most prestigious French engineering and business schools, and will be the largest French university in the new national organization of academic structures.

CDS gathers 250+ permanent researchers. The project is funded for 2 years, with a probable 3rd year extension, and is supposed to evolve towards a permanent structure, in relation with the strategic plans for scientific computational environments currently developed by the French agency for scientific research (CNRS) and Ministry of Higher Education and Research. The budget is 1.2 M€ for 2 years.

CDS has a specific task, called interface to data centers, that specifically targets a data platform for the collaboration.

The summary of the overall project is as follows:
The subject of data science is the design of automated methods to analyze massive and complex data in order to extract useful information from them. Data science projects require expertise from a vast spectrum of scientific fields ranging from research on methods (statistics, signal processing, machine learning, data mining, data visualization) through software building and maintenance to the mastery of the scientific domain where the data originate from. The goal of this initiative is to establish an institutionalized agora in which these scientists can find each other, exchange ideas, initiate and nurture interdisciplinary projects, and share their experience on past data science projects. To foster synergy between data analysts and data producers we propose to provide initial resources for helping collaborations to get off the ground, to mitigate the non-negligible risk taken by researchers venturing into interdisciplinary data science projects, and to encourage the use of unconventional forms of information transmission and dissemination essential in this communication-intensive research area (such as brainstorming sessions or data challenges). The CDS would fit perfectly in the recent surge of similar initiatives, both at the international and at the institutional level, and it would make the University one of the international forerunners of data science. The CDS will naturally coexist and collaborate with existing structures, including six Labexes, doctoral schools, and M.Sc. programs. Although the primary focus of the initiative will be on scientific data, it will also be in a perfect position to play the role of a contact point to industrial partners.

TOC
Many disciplines are covered. The relevant part of the table of contents of the project is as follows :

Q3: What are the scientific disciplines covered by your research activities? (select one or more from the EGI scientific discipline classification)

Discipline 1	Many disciplines are covered, see the table of contents in the previous section

Q4: Data products: what are the data sources, what type of data formats and what is the amount of data produced including the frequency of generation/update?

The CDS has been created very recently (30 January 2014). The “interface to data centers” task will map the (numerous) resources of the partners. As such, the answers to questions 4-10 are replaced by the description of this task.
SUMMARY. For a long time, the national and local research institutions have deployed pooled (shared) IT infrastructures that enable sharing physical resources and expertise, creating key capabilities and ex- perience in this process. A paradigm shift occurs with the current call of national research institutions for an integrated approach, with interdisciplinary structures embedded in the local environment; the structure is organized around a permanent high-level and open technological platform operated and animated by qualified staff. In our case, Data Science is based on Information Technology (IT). A Data Science IT Platform is the component of the IT system that should provide an effective and efficient platform for empowering the communities of researchers to create and share knowledge: develop, eval- uate, exploit, and disseminate software, build and share new applications, share access to common data repositories, and create information sources and new services. The CDS will contribute to the definition, organization, and implementation of a Data Science IT Platform within UPSa at two levels: the CDS can be used by the governing bodies as a transversal entry point for organizational and infrastructure ac- tions; and the CDS will develop prospective studies and proof of concept implementations for advanced sharing and knowledge building.
CONTEXT. The pooling of physical, immaterial, and human IT resources is driven by multiple factors. The creation of large, energy-efficient data centers is motivated by the changing structure of the opera- tional costs that stems from technology trends. Software

efficient data centers is motivated by the changing structure of the operational costs that stems from technology trends. Software sharing is motivated by the need for avoiding redundant activities between science fields or between science fields and industry and preventing the technological balkanizations that precludes interoperability. Data sharing, beyond its general necessity to simply perform state-of-the-art research, is the first step to build the critical mass required for main- taining research data for reuse and preservation through its lifecycle of interest. Finally, the new frontier of IT for research is automated support for knowledge extraction and sharing, with a growing concern for reproducibility.

ASSETS. The CDS collaboration features exceptional assets to advance a Data Science IT Platform at UPSa.

Large scale infrastructures VirtualData isa data center housing infrastructure designed for high energy efficiency. The resources are exploited mainly under a grid or a cloud (StratusLab) model. Virtual- Data has been initiated by eight physics laboratories within the P2IO Labex; it pools 6000 cores and 3PB of disk,on two locations(University Paris Sud and EcolePolytechnique). University of Paris Sud is building a roadmap to exploit the housing infrastructure (with or without cloud technology) and more extensive resource pooling. IDRIS is one the High Performance Computing (HPC) national centers, and a major resource of competence for HPC activities, with a particular interest for us due to its historic interactions with other national and local facilities (Maison de la simulation, Meso-centre Ecole centrale) in the campus geographic area. BADAP is a project jointly led by the GENES and the IMT—funded by the program "investissements d'avenir"—of a Big Data platform oriented towards research and innovation. Besides several hundreds of terabytes of storage facilities and 4TB of memory, it also offers a small team of qualified specialists for accompa- nying scientific projects related to Big Data along with a technology (through the CASD) providing secure access to sensitive or confidential data.

Innovative exploitation A Data Science IT Platform calls for a systemic approach that exploits synergy among computer research communities who see it as an object of research, and other research communities who see it as a platform in service of research. A long-lasting scientific interdisciplinary collaboration already exists in this domain with the Grid Observatory project, as well as relevant basic research (see Section 3.9).

Reproducible research CMLA has promoted the foundation of a new kind of scientific journal, whose first working example is the prototypical journal Image Processing On Line (IPOL) founded in 2010. IPOL publishes peer-reviewed papers describing the algorithms in accurate literary form, coupled with code. Furthermore it allows scientists to check directly the published algorithms on line by providing a web execution interface on any uploaded image. An archive associated with each article permits researchers to share their experiments. Finally, data papers are encouraged, where researchers can publish papers linked to databases that they wish to share. The acquisition must be described accurately in the IPOL paper. With a growing experimental archive of more than 100000 original image data used online by researchers worldwide, IPOL demonstrates the efficiency and appeal of online execution to foster reproducible research and interdisciplinary communication. CMLA can therefore bring to the CDS its experience, technique and software in the online publication of shared data and software, with a set up that rewards researchers doing this effort.

A data-intensive scientific environment Most participants of CDS are involved in a rich ecosystem of data-oriented scientific environments, through thematic networks, Equipex facilities, and local data acquisition facilities. In many cases, they are involved in the policy design for scientific IT at their respective institutions. These positions will contribute to a short feedback loop between CDS and the relevant institutions, and contribute to a better integration of data-intensive IT at the UPSa level.

OBJECTIVES. The functionalities of a Data Science IT Platform can be organized along the 1.0/2.0/3.0 terminology: even if the underlying technologies and goals are not essentially web-oriented, the final goal is the same: make the experience of the user, in our case the research activity of data scientists, much easier and more productive.

1.0: Raising awareness. We will assess the basic requirements (computational power and storage) of the CDS partners in the short (project duration) and medium (5 years) term. We will also map the existing data repositories operated by CDS partners that are relevant for the interdisciplinary research of CDS, and their relations to wider scientific networks. The expected results in general are a better knowledge of resources and needs within the project, and promotion of pooling ex- periments, including cloud technologies and parallel systems. Specific outcomes can be expected both at the disciplinary level - new sharing of data, infrastructure or expertise - and at the inter- disciplinary level, with a particular focus on making available truly large datasets for algorithmic research internally.

2.0: A collaborative interface. Scientists should be able to do more than just retrieve information, by interacting and collaborating as creators of user-generated content in a virtual community. At this level, the content should probably be limited to software and data; needless to say, to the extent that the researchers accept to share them. The essential envisioned capabilities are firstly to make these software and data actually usable by a human researcher, and second to facilitate the related interactions: information should be richer, easier to find and more thoroughly categorized than by the usual static "portfolios". The expected results are threefold: firstly, actual proof of concepts developments for cross-exploitation of codes and data, organized around a shared infrastructure; second, sharing of programming expertise; and finally a structured methodology derived from theses experiences.

3.0: The computer generates new information. The goal is to create a capability that provides seamless access to effective and personalized science aids. Within this wide and still speculative area, the CDS will explore two precise avenues. The first one is the Linked Data framework, by exploring how related data repositories operated by CDS partners could be seamlessly and meaningfully queried with the specific objectives of Data Science. The second is reproducibility/repeatability of experiments. This subject is receiving increasing attention; elaborating on the interaction of the existing exceptional experience at CMLA, the scalable cloud infrastructure at VirtualData, and the scientific expertise of CDS data providers and analysts would be a powerful instrument of cohesion for the project, and offer the perspective of a "killer application".

ACTIONS.These goals will be implemented through the following activities.

Web portal Within the CDS general portal (action interactive web portal), a portal dedicated to the imple- mentations of the above-mentioned goals will be created. The envisioned timeline is T0+6 months for the 1.0 goal, T0+12 and T0+24 for experiments on respectively the 2.0 and 3.0 goals.

Proof of concept developments Cross-exploitation of codes and data as well as the deeper integrations of 3.0 objectives are likely to raise two kind of adaptation issues: interoperability at the applicative level, and accessibility at the infrastructure level. The specific role of this action will not be to provide support for the strictly applicative developments that, at this stage, will better be realized directly within the research teams (possibly through CDS funding, either coding sprints or data challenges actions), but to advise on good practices and implement the tools required to make them sustainable. Accessibility encompasses the obstacles that can be specifically encountered with shared or cloud infrastructures, and should be supported by the coding sprints action.

Roadmap This activity targets the contribution of CDS to the wider UPSa initiative for organizing its research computing infrastructures. The WG will provide the scientific interface to the Data Science community by its feedback along the project, and by contributing to a roadmap for the Data Science IT within UPSa.

The technical developments will be realized by a research engineer, with strong skills in semantic data organization and collaborative technologies, for the duration of the project.

Q5: Data Storage: what digital data is retained during the life of the project/experiment? In what media and location?	*Respondent skipped this question*
Q6: Data Management Plans: Do you have a well defined data management plan?	*Respondent skipped this question*
Q7: Meta data collection, management, and packaging: Please indicate, how you collect, manage, and package your meta data in order to prepare it for supporting archival and preservation of your research data and what type of other meta data you (do/would like to) store along with your research data. What metadata standards do you currently use? Please describe the technical challenges to be addressed.	*Respondent skipped this question*
Q8: Persistent Identifier: Please indicate, how you generate your persistent identifiers for your data?	*Respondent skipped this question*
Q9: Preserved data: Please indicate the type(s) of data and the amount of data (GB per year, overall)	*Respondent skipped this question*
Q10: Support for access and re-use: How do you presently store your digital research data for future access and use? What kinds of future re-use cases could you envisage? How much of your data do you make available to others? (none, some, most, all) If all or part of your data are not available to others, why not? What type of access policy is in place (e.g., open access)? What type of security is implemented? (e.g., open, local password, LDAP, X.509 certificate)	*Respondent skipped this question*
Q11: Business models: What is the current business model for the gathering, archival, preservation, and re-use of the data?	*Respondent skipped this question*

Q12: **Can you describe if/what educational and training needs you could identify in any of the areas mentioned above? This will very important for us so to define future educational and training programs.**

The CDS has an educational action. Its goal is to catalyze the data science curriculum of UPSay

Q13: **Would you in general be willing to present a use case/scenario/requirement/demand from the above in the workshop**

We can present an overview of CDS and its data related plans, with a focus on the needs of scientific research, including advanced data representations and experiment repeatability. The goal of our contribution is to explore the possibility of international collaboration with similar projects.

#8

COMPLETE

Collector: Web Link (Web Link)
Started: Monday, February 24, 2014 9:42:43 AM
Last Modified: Monday, February 24, 2014 2:31:26 PM
Time Spent: 04:48:43
IP Address: 2.227.191.236

PAGE 1

Q1: **Contact Information**

Name:	Antonella Fresa
Organisation:	Promoter Srl
City/Town:	Peccioli (Pisa)
Country:	Pisa
Email Address:	fresa@promoter.it

Q2: What is the research group or project for which you are providing information about research data lifecycle? Please, provide some background information on the size and geographical distribution of the research collaboration/project and about the main scientific challenges being addressed.

We are providing information related to the Digital Cultural Heritage (DCH) sector.

The volume of digital cultural heritage data is incredibly growing year after year, so now it is necessary to reflect upon the tools which permit to manage such a huge amount of data in an efficient and selective way, in order to make the data available to the researchers and the citizens in a European dimension.

The needs of DCH sector are:
• high quality information technology management, to ensure trust, availability, reliability, long-term safety of content, security, preservation and sustainability;
• enhanced access facilities
o to the researchers who will look for contents into the DCH e-Infrastructure for their research;
o to the cultural institutions that will deliver their data to the DCH e-Infrastructure;
• interoperation among existing cultural heritage repositories, among cultural portals and among data from the digital cultural heritage and from the research.

Main challenges are:
• High investment for in the production of DCH data due to the need of human intervention of experts.
• High costs of digital preservation, due to the use of separate solutions implemented by each memory institution.
o The estimated total cost of digitising the collections of Europe's museums, archives and libraries, including the audiovisual material they hold is approximately €100bn, or €10bn per annum for the next 10 years
o The cost of preserving and providing access to this material over a 10-year period after digitisation would be in the order of €10bn to €25 bn, provided that centralised repository infrastructure is made available for the purpose
• DCH content is difficult to be preserved because data are complex and interlinked through many relations.
• Contextual data are very important for cultural research.
• The digitisation process is unique cannot be replicated unless the whole work is done from scratch.

2 twin-projects (DC-NET and INDICATE) and an ongoing international coordination action (DCH-RP: Digital Cultural Heritage – Roadmap for Preservation) brought together in the last years memory institutions and e-infrastructure providers from all over Europe to work for the future, in order to create a data infrastructure devoted to cultural heritage research. In particular, long-term preservation of digital cultural content has been identified as the highest priority for the DCH sector.
The potential benefits from the use of e-infrastructure are:
• To allow for cost reduction in digitisation, cataloguing and metadata generation by substituting expensive human workforce with cheaper machine processes
• To support the permanent identification of digital cultural objects and providers
• To facilitate storage and preservation, ranging from short- medium- and long-term
• To improve search facilities to manage semantic search and linked open data
• To enhance processing and visualisation of complex cultural data (e.g. 3D modelling and VR representations) through the computing resources offered by research e-infrastructures (grid, cloud)
• To enable dynamic distributed virtual organisations, facilitating collaboration with information and resource sharing (e.g. virtual conferences, document sharing, blog and cooperation platforms, etc.)
• To contribute to standardisation in the data world, e.g. by developing a common reference model for the DCH sector

These initiatives are contributing to smooth the way to the Open Science Infrastructure for Digital Cultural Heritage which is foreseen in 2020.

Other valuable initiatives in this field are:
• PREFORMA (PREservation FORMAts for culture information/e-archives) a Pre-Commercial Procurement (PCP) co-funded by the European Commission under its FP7-ICT Programme to address the challenge of implementing good quality standardised file formats for preserving data content in the long term. The main objective is to give memory institutions full control of the process of the conformity tests of files to be ingested into archives.
• Europeana Photography, a Pilot B action putting together and sending to Europeana some of the most prestigious photographic archives, public libraries and photographic museums covering specifically the length of time from the beginning of photography (1839 with the first example of images from Fox Talbot and Daguerre) to the beginning of the Second World War (1939). Special attention is devoted to the management of intellectual property, which is further emphasised by the involvement of content providers from both the private and the public sectors.
• Europeana Space, a Best Practice Network with the aim to increase and enhance the creative industries' use of Europeana by delivering a range of resources to support their engagement. Europeana Space will address all sectors of the creative industries, from content providers to producers, exhibitors, artists and makers of cultural/creative content, publishers, broadcasters, telecoms and distributors of digital content. The project aims to address the problems, which limit the re-use of Europeana by the creative industries, such as issues around the IPR status of content and the need for business models demonstrating the potential for exploitation of available content.

Q3: What are the scientific disciplines covered by your research activities? (select one or more from the EGI scientific discipline classification)

Discipline 1	History and Archaeology
Discipline 2	Arts

Q4: Data products: what are the data sources, what type of data formats and what is the amount of data produced including the frequency of generation/update?

The European amount of digitized material is growing very rapidly, as National, regional and European programmes support the digitization processes by Museums, Libraries, Archives, Archaeological sites and Audiovisual repositories.
Recent studies commissioned by the EC (NUMERIC Study Report: http://cordis.europa.eu/fp7/ict/telearn-digicult/numeric-study_en.pdf; ENUMERATE Survey Report on Digitisation in European CH Institutions 2012: http://www.enumerate.eu/fileadmin/ENUMERATE/documents/ENUMERATE-Digitisation-Survey-2012.pdf; EC Comité des Sages Report on Cost of Digitising Europe's CH: http://ec.europa.eu/information_society/activities/digital_libraries/doc/refgroup/annexes/digiti_report.pdf) showed that:
• 83% of institutions said curatorial care is part of the mission
• 83% of institutions have a digital collection or is currently involved in digitisation activities
• 20% of all collections, that need to be, are digitised
• 89% of audio visual institutions have born digital collections, while 43% of museums of art and history have them
• 34% of institutions have a written digitisation strategy
• About one third of the institutions are included in a national digitisation strategy, for national libraries more than half are included

DCH content is composed of several different kind of information and different formats: texts, still images, 3D models, publications, digital exhibitions, virtual reconstructions, etc.

Examples of standardised formats often used by memory institutions are:
• Document formats. Public authorities and other institutions producing electronic documents and media content on national level are normally using open standards adapted to specific requirements to produce their electronic files. PDF/A, and its different versions, is for example the standard mostly used by archiving institutions for electronic documents.
• Image formats. TIFF is the preservation format most often used by memory institutions for still image digitisation.
• Audiovisual formats. The Material eXchange Format (MXF) is a container format for professional digital video and audio media which is developed and maintained by audio-visual industry, particular for postproduction and distribution purposes. In the near future, memory institutions will have to deal with very large numbers of these files that will be produced by software that claims to support the specific format.

Q5: Data Storage: what digital data is retained during the life of the project/experiment? In what media and location?

In the DCH sector, data which are digitised are then retained. The main retention requirements are:
• Separation of content and metadata;
• OASIS compliance;
• Accessibility through retrieval and search system.

Usually, data and metadata are stored in data centres / repositories hosted by the memory institutions themselves, even if this poses big maintenance issues due to the lack of ICT expertise.
Studies conducted in the projects mentioned above showed that access to shared resources, as in the case of the e-infrastructures (cloud, grid), is very appealing. The main problems in adopting this approach are:
• Issues related to copyrights.
• Cultural data are curated by many different persons: data management and administration + user access control are very important.
• Security of the data is very important for cultural institutions: trust building is a key factor when it is not determined where data are stored.
• Functionalities and services offered by e-infrastructures should not impact on the outgoing traffic of the institution.
• Access to the e-infrastructure services should be simple without requiring IT specialist knowledge.

Q6: Data Management Plans: Do you have a well defined data management plan?

No,

If yes, add the URL
CH institutions often do not have a well defined data management plan. Each country, each sector and often each organisation have different policies and guidelines for accessing, sharing and processing the content under its control.

Q7: Meta data collection, management, and packaging: Please indicate, how you collect, manage, and package your meta data in order to prepare it for supporting archival and preservation of your research data and what type of other meta data you (do/would like to) store along with your research data. What metadata standards do you currently use? Please describe the technical challenges to be addressed.

The extensive use of relevant and open standards is a vital pre-requisite for the CH community to promote interoperability, encourage widespread access and control costs in its digital preservation programmes.
Extensive reviews under the auspices of the Minerva (2008), Athena (2009), Linked Heritage (2011)and DCH-RP (2013) projects categorized and described many of the standards that are most applicable or recommended in this area. The more relevant deliverables from the earlier projects are available as follows:
• Athena: http://www.athenaeurope.org/index.php?en/149/athena-deliverables-and-documents
o D3.1, Report on Existing Standards Applied by European Museums.
o D3.2, Recommendations and Best Practice Report.
• Linked Heritage: http://www.linkedheritage.eu/index.php?en/142/documents-and-deliverables
o D2.1, Best practice report on cultural heritage linked data and metadata standards.
o D2.2, State of the art report on persistent identifier standards and management tools.
• DCH-RP: http://www.dch-rp.eu/index.php?en/61/deliverables

o D3.2, Standards and interoperability best practice report.

The following list summarizes a number of important standards in the CH sector, together with several standards such as LIDO that are of a more general or cross-cutting nature.

EAD
Encoded Archival Description
Archive
An XML standard for encoding archival finding aids
http://www.loc.gov/ead/ead.xsd (W3C schema)

ISAD (G)
General International Standard Archival Description
Archive
General rules for archival description that may be applied irrespective of the form or medium of the archival material
http://www.icacds.org.uk/eng/ISAD(G).pdf

OAIS
Open Archival Information System, ISO 14721:2012
Archive, cross-domain
A conceptual reference model for an open archival information system (OAIS). An OAIS is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a “designated community”
http://public.ccsds.org/publications/archive/650x0m2.pdf

MoReq2
Model Requirements Specification for the Management of Electronic Records
Archive, electronic records
An XML standard for electronic records, developed by the DLM Forum
http://www.moreq2.eu/home

ISAAR (CPF)
International Standard Archival Authority Record for Corporate Bodies, Persons and Families
Archive, party information
A standard for authority records for corporate organizations, persons and families
http://www.icacds.org.uk/eng/ISAAR(CPF)2ed.pdf

ONIX for Books
Book publishing and supply
An XML-based standard for communicating book product information
http://www.editeur.org/83/Overview/

VCAR
Business, party information
A standard for describing electronic business cards
http://tools.ietf.org/html/rfc6350

Indecs
Interoperability of data in e-commerce systems
Conceptual, e-commerce
Provides a framework for metadata requirements for e-commerce in content (intellectual property), focusing on semantic interoperability
http://www.doi.org/topics/indecs/indecs_framework_2000.pdf

DBpedia Ontology
Cross-domain
A cross-domain ontology, based on the “infoboxes” of Wikipedia
http://dbpedia.org/Ontology

Europeana Data Model
Cross-domain
Created for structuring data for Europeana ingestion, management and publication, and improves on Europeana’s basic data model, the Europeana Semantic Elements (ESE)
http://pro.europeana.eu/edm-documentation

VRA Core
Visual Resources Association
Cross-domain
A data standard for the description of works of visual culture and the images that document them
http://www.vraweb.org/projects/vracore4/index.html

Premis
Preservation Metadata: Implementation Strategies
General heritage
An XML-based metadata framework
http://www.loc.gov/standards/premis/

CIDOC-CRM
CIDOC Conceptual Reference Model
General heritage, conceptual
Provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation
http://cidoc.ics.forth.gr (CRM website)

Dublin Core
General, cross-domain
A simple metadata element set intended to facilitate discovery of electronic resources
http://dublincore.org/documents/dc-rdf

SKOS
Simple Knowledge Organization System
General, cross-domain
Designed for the publication of controlled structured vocabularies for the Semantic Web, including thesauri, classification schemes, taxonomies, and subject headings
http://www.w3.org/TR/2009/REC-skos-reference-20090818

Basic Geo Geography
An RDF vocabulary for basic geographical information: latitude, longitude, and altitude
http://www.w3.org/2003/01/geo

MIDAS Heritage
Historic environment
A standard for the management of the historic environment
http://www.english-heritage.org.uk/publications/midas-heritage/midas-heritage-2012-v1_1.pdf

FRBR
Functional Requirements for Bibliographic Records
Library
A conceptual entity-relationship model for use with online library catalogues and bibliographic databases
http://www.ifla.org/VII/s13/frbr/frbr.pdf

MAB2
Maschinelles Austauschformat für Bibliotheken
Library
An automated exchange format for libraries, since superceded by MARC-21
http://www.dnb.de/EN/Standardisierung/Formate/MAB/mab_node.html

MARC
MAchine Readable Cataloging
Library
A set of digital formats for the description of items catalogued by libraries, such as books
http://www.marc21.ca/index-e.html (MARC21)
Five types of formats are supported, respectively for Authority (or authorized form), Bibliographic, Classification, Community Information and Holdings
http://www.openarchives.org/OAI/2.0/guidelines-marcxml.htm (MARCXML)

METS
Metadata Encoding and Transmission Standard
Library
An XML-based standard for encoding descriptive, administrative and structural metadata regarding objects within a digital library
http://www.loc.gov/standards/mets/

MODS
Metadata Object Description Schema
Library
An XML-based bibliographic description schema
http://www.loc.gov/standards/mods/

CDWA
Categories for the Description of Works of Art
Museum
Describes the content of art databases by articulating a conceptual framework for describing and accessing information about objects and images
http://www.getty.edu/research/institute/standards/cdwa/index.html

museumdat
Museum
A harvesting format for providing core data from museum holdings
http://museum.zib.de/museumdat/museumdat-v1.0.xsd (XML schema)

Object ID
Museum

Description of cultural objects, especially of use if objects are stolen
http://archives.icom.museum/objectid/

SPECTRUM
Museum
A collections management standard including data elements for management and description
http://www.collectionslink.org.uk/spectrum-standard

LIDO
Lightweight Information Describing Objects
Museum, general
A harvesting format for collections, particularly from museums, to portals
http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf

MO
Musical Ontology
Music
Contains concepts and properties for describing music, for example: Artists, Tracks, Performances, Arrangements
http://musicontology.com

MusicXML
Music
An XML-based file format for representing Western musical notation
http://www.musicxml.com/

FOAF
Friend Of A Friend
Party information, roles
A format, using RDF and OWL, for describing persons, their relations to other persons and things, and their activities
http://xmlns.com/foaf/spec

BIBO
Bibliographic Ontology
Publications
A format, using RDF, for describing bibliographic items like books, magazines, and newspaper pages
http://bibliontology.com/specification

Furthermore, the OAI-PHM/OAI-DC standards are used in certain cases to aggregate data and to make data available for publication in other portals (as in the case of Europeana).

Q8: Persistent Identifier: Please indicate, how you generate your persistent identifiers for your data?

The PID requirements do not vary significantly from one DCH initiative to another – this represents a service useful to most DCH work, 'out of the box'. There is excellent research and development already done in this area, and one or more of the existing schemes (e.g. DOI, ARK, URI, URN, etc.) could be adopted with minimal adjustment.

This topic was covered at some length within Deliverable D2.2 of the Linked Heritage project. The report explained the role and importance of such identifiers before describing the primary candidate identifier types for use in the Cultural Heritage arena. Fundamental features of three general digital identifier standards – URI (Universal Resource Identifier), URL (Universal Resource Locator) and URN (Universal Resource Name) – were presented first and they are included in the table in section 4.1.5 above. The report then described four types of service-associated digital identifier standards: PURL ((Persistent URL) & Handle System), DOI (Digital Object Identifier), OpenURL and ARK (Archival Resource Key). Summary information on these four is presented in the list below.

ARK
Archival Resource Key
A URL scheme which can identify both physical and digital objects
http://en.wikipedia.org/wiki/Archival_Resource_Key

DOI
Digital Object Identifier
A stored and maintained character string used to uniquely identify any kind of entity, physical, digital or abstract
http://www.doi.org/
The DOI is associated with a prescribed set of metadata. Used in conjunction with the Handle system, the DOI provides an infrastructure for the persistent identification and location of digital resources

OpenURL
A URL with embedded metadata (used by resolver services) to more easily find a resource
http://en.wikipedia.org/wiki/OpenURL

PURL
Persistent URL
A URL pointing to a resolver (e.g., a handle) which directs to a current URL
http://purl.oclc.org/docs/help.html#overview

Arguably, the service-associated and maintained identifiers are likely to offer more comprehensive features to CH institutions managing digitized resources, but issues relating to both cost and policy have militated against the widespread adoption of such identifiers in this area.
The example of DOIs may be illustrative here. Although the utility of this identifier system is clear, uptake to date in the DCH environment has been very low. The costs associated with DOI registration may represent one significant barrier. But equally, the absence of a DOI Registration Agency specifically aligned with the DCH community – understanding cultural heritage requirements and able to design registration metadata that is meaningful to that community – may represent an even greater impediment. (Compare for instance the near-ubiquity of DOIs in the area of scientific, technical and medical journals, which are serviced by the Crossref registration agency, itself originally established by the STM publishing industry itself.)

Q9: Preserved data: Please indicate the type(s) of data and the amount of data (GB per year, overall)

In general:
• 23% of institutions have a written digital preservation strategy, figures range from 44% for national libraries to 12-25% for museums
• About a third of the institutions are included in a national preservation strategy
• 40% of national libraries say there is no national digital preservation strategy
• 30% of the institutions are included in a national digital preservation infrastructure

Type and size vary from case to case.
Types include: texts, still images, 3D models, publications, digital exhibitions, virtual reconstructions, etc.
Size range from 5 to 200 GB.

Q10: Support for access and re-use: How do you presently store your digital research data for future access and use? What kinds of future re-use cases could you envisage? How much of your data do you make available to others? (none, some, most, all) If all or part of your data are not available to others, why not? What type of access policy is in place (e.g., open access)? What type of security is implemented? (e.g., open, local password, LDAP, X.509 certificate)

The creation of an online "location" for the presentation of DCH materials online is a central part of any digital heritage initiative. The model applied will most commonly be a content management system, a portal system, a digital library or digital repository which has been specifically designed and built for the purpose.

Generally, most of the data are available to others, apart from those that are not totally documented.
Usually, access to these data is open for view only purposes and password protected for importing and updating data.
Adding and editing data needs to be password protected and limited to known individuals authorised by the institution.
Authentication mechanisms most in use: Open access, Password protected, IP-based, X.509 certificate based, Shibboleth or equivalent.

Concerning re-use
• 31% of the institutions have a policy on the use of the digital collections, figures range from 60% for national libraries to 22% for archaeology museums
• 42% of institutions monitor the use of their digital collection
• 85% of institutions use web statistics to measure the use of their digital collections
• By 2014 institutions estimate to make twice as much of their collections accessible through Europeana when compared to today

The end users of DCH work tend to be DCH researchers and/or members of the general public.
Other use cases include re-use of data in education, in commercial ventures, in collaborative projects and for digital exhibitions.

Finally, in the last years, Cultural and Creative Industries revealed to be another interesting possibility to exploit the potentials of digital cultural content as an essential driver of creativity, innovation and competitiveness in the framework of a sustainable economy, as highlighted first by UNESCO and then by the European Commission.
Cultural Industries comprise the economic domains that focus on Cultural Heritage, museums and libraries, cultural and heritage tourism and also education as well as research in cultural domains. Creative Industries comprise domains such as arts (visual and performing arts) and architecture, design, crafts, fashion, music, film, publishing, advertising, TV and Radio, toys, video games and serious-educational games, software, as well as research and development in technological domains (R&D).

Q11: Business models: What is the current business model for the gathering, archival, preservation, and re-use of the data?

Memory Institutions (museums, archives and libraries at first, but also Archaeological sites and Audiovisual repositories) need to digitize their content both for preserving it in a digital format and for granting and enlarging the access to them by researchers, students and citizens. It is esteemed that only a very small part of the European cultural heritage had been digitized until now, therefore there is a lot of work to do and memory institutions are bearing big efforts to carry on this huge deal. A growing number of projects for digitization is supporting the process and indeed the cultural heritage sector is going through wide transformations and changes.
The problem is that memory institutions' budgets often don't meet the requirements of such projects. In search of funding issue solution cultural heritage institutions tried many variants – lobbying activities to establish additional governmental funding, participation in the projects of international organizations, partnership with other memory institutions to share the costs of digitization, seeking sponsor help and so on.

Commercial promotion of cultural heritage can be nowadays a valid financial source, in particular in the education and tourism (cultural tourism) sectors and by the Creative Industries. The use of cultural heritage content by the creative industries is still limited by factors including issues around the IPR status of content and the need for business models demonstrating the potential for exploitation of available content. But there are projects that have been funded in the last years by the European Commission to address these issues (such as Europeana Space). The aim is to experiment with innovative applications and services the creative re-use of cultural resources, to allow cultural institutions to become both content providers and service providers, exploring new audiences and markets for cultural heritage bodies and promoting further investment in digitisation of cultural content.

In this context, public-private partnerships have also been taken into account as possible ways to exploit the digital resources made available by the memory institutions, as demonstrated by the establishment of dedicated task forces in Europeana and in other related projects (such as Linked Heritage and Europeana Photography).

The embracing of the e-Infrastructures by the digital cultural heritage community can open new scenarios of use and exploitation with an impact expected on different sectors:
• The cultural heritage sector. The managers who work in the Cultural sector can become more aware about the potential that the e-infrastructures can offer to their work: storage, preservation, access services for the cultural institutions, etc.
• The research. A better integration of the cultural sector with the e-Infrastructures can enable the research of new advanced services and applications.
• The economic sectors. Digital cultural content can become more usable and re-usable for education, cultural tourism, long-life learning, non-professional cultural interests, creative industry, etc.

Q12: Can you describe if/what educational and training needs you could identify in any of the areas mentioned above? This will very important for us so to define future educational and training programs.

• ICT, metadata architectures, data modelling, DB management, web engineering, new media management, data warehouses and repositories, W3C/ISO standards and technologies, interoperability/access protocols, data mining, filtering
• Digitization techniques
• Intellectual Property Management
• Storage, archiving, preserving methods and strategies
• Information design, web design, usability engineering, user studies, web services, CRM, communication strategies
• Digital library management, collection management, collaborative/federated approaches, content management, project mgmt., cross-cultural and multilingual issues
• Knowledge organization, ontologies, advanced indexing and retrieval methods, visualization
• High technology illustrations such as 3D modelling, augmented reality, etc.

Q13: Would you in general be willing to present a use case/scenario/requirement/demand from the above in the workshop

Yes.