Data Management Plan

Authors: Hannu Saarenmaa (UH), Donat Agosti (Plazi), Simon Chagnoux (MNHN), Emily van Egmond (Naturalis), Quenting Groom (APM) & Alex Hardisty (CU)

DELIVERABLE REPORT

Grant Agreement Number: 777483 / **Acronym:** ICEDIG

Name of project: Innovation and consolidation for large-scale digitisation of natural heritage

Start date: 01 Jan 2018 / **Duration:** 27 months

Call: H2020-INFRADEV-2017-1 / **Type of Action:** RIA

1. Deliverable: **D6.1**
2. Deliverable name: **Data Management Plan**
3. Work package number: **6**
4. Short name of the Lead Beneficiary: **UH**
5. Type: **Report**
6. Dissemination level: **Public**
7. Delivery date: **Month 3, 2018-03-31**

Executive summary

ICEDIG is a design study for the proposed new research infrastructure *Distributed System of Scientific Collections* (DiSSCo), focusing on the issues around digitisation of the collections and making their data freely and openly available following the FAIR principles (data being Findable, Accessible, Interoperable, and Re-usable).

As a design study, ICEDIG does not implement anything in an operational fashion, and therefore the amount of research data which ICEDIG will deal with is limited. DiSSCo, on the other hand, is expected to deal with huge amounts of data. The data management plan (DMP) for DiSSCo will be produced as one of the design documents of ICEDIG. In other words, ICEDIG will produce two data management plans: one for ICEDIG and another for DiSSCo.

In order to achieve its objectives, ICEDIG will carry out a number of tasks, including specific pilots which may produce limited amounts of research data. ICEDIG will also deploy large volumes of already existing research data in order to explore how it can best be pooled in the available European Open Science Cloud infrastructures.

This data management plan follows the Horizon 2020 DMP template, which has been designed to be applicable to any Horizon 2020 project that produces, collects or processes research data. The template is a set of questions that have been answered with a level of detail appropriate to the project. It has been understood that the DMP is a living document which may be updates as the implementation of the project progresses and when significant changes occur. Therefore, the DMP has a clear version number and includes below a timetable for planned updates.

Table 1. Timetable for updates of the Data Management Plans of ICEDIG and DiSSCo.

Version	Date	Changes
Draft versions v0.1 and v0.2	2018-03-15/26	ICEDIG DMP created
Final v1.0	2018-03-31 (Month 3)	ICEDIG DMP submitted as D6.1
Final v2.0 revision	2019-03-31 (Month 15)	To reflect experiences with data storage tasks and other prject findings
Final v3.0 ICEDIG D6.5 to be submitted	2019-09-30 (Month 21)	ICEDIG DMP to be revised with findings from the deliverable report *Open access implementation guidelines for DiSSCo*
Final v4.0 DiSSCo DMP to be submitted as ICEDIG D6.6	2019-10-31 (Month 22)	ICEDIG DMP to be revised with findings from the deliverable report *Provisional Data Management Plan for the DiSSCo infrastructure*

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

As a Design Study of the Infrastructures work programme of H2020, the general objective of the project ICEDIG is to lay the ground work for the proposed ESFRI Research Infrastructure *Distributed System of Scientific Collections* (DiSSCo), firstly by supporting the technological innovations that will be needed to efficiently digitise one and a half billion collection objects in a foreseeable time, such as the next 30 years, and secondly to consolidate the organisation that needs to perform this task. When this has been achieved, the natural science community will be a fully enabled player in digital society, and the most fundamental scientific data on the diversity of life on earth will be freely and openly available for all.

The specific objectives of the project ICEDIG are to:

- Determine the technological and logistical basis for extracting images and other relevant data from tens of millions of physical objects annually, across Europe;
- Determine how to make this information discoverable through efficiently capturing the essential metadata of each object;
- Create plans for storing that huge information pool efficiently in national and European open science infrastructures;
- Involve researchers, citizen scientists, industry, and the society at large in this huge effort;
- Summarise the above knowledge in business and engineering plans for the development of the new RI that is DiSSCo.

ICEDIG is a Design Study for DiSSCo, focusing on the issues around digitisation. ICEDIG does not implement anything in an operational fashion, and therefore the amount of research data which ICEDIG will deal with is limited. DiSSCo, on the other hand, is expected to deal with huge amounts of data. The *data management plan* (DMP) for DiSSCo will be produced as one of the design documents of ICEDIG. In other words, ICEDIG will produce two data management plans: one for ICEDIG and another for DiSSCo. This document is the former one, and due at Month 3 of the project as the Deliverable D6.1. A “Provisional Data Management Plan for the DiSSCo infrastructure” is the Deliverable D6.6 of the ICEDIG project, and due at Month 22 of the project.

In order to achieve its objectives, ICEDIG will carry out a number of tasks, including specific pilots which may produce limited amounts of research data. ICEDIG will also deploy large volumes of already existing research data in order to explore how it can best be pooled in the available European Open Science Cloud (OSC) infrastructures.

What types and formats of data will the project generate/collect?

In general, ICEDIG will deal with data that can be generated by digitising the objects in scientific collections. These objects include herbarium sheets, pinned insects, bones and skins of vertebrates, liquid samples of fish, reptiles and invertebrates, microscopic slides, etc. In addition, geological and paleontological samples should be dealt with. Scientific collections also are curators of field notebooks, photographs, paintings, etc., which also to be digitised as well.

Digitising these objects means extraction of the features of the objects, i.e., taking photographs and measures of the objects, and extracting DNA and chemical samples. It also means entering metadata of the collecting events of the samples, including location, time, methods, collecting agent, etc. Taxonomic

identification of the organisms contained in the samples is a major activity that must be performed, often repeatedly when our knowledge of the taxonomic classification accumulates.

By digitising these objects, digital surrogates of the physical objects can be created. This does not mean replacing the physical object, but only allows wider access to the information about the physical object. If packaged appropriately as actionable digital objects, the management of the diverse data can be streamlined. This will be a major topic in the DiSSCo DMP.

In its work package (WP) tasks and pilot projects, ICEDIG will deal with limited amounts of such data, which has been summarised in Table 1 below. They may be generated in the following actions:

- **Technology pilot projects in WP3.** These belong to the subtask T3.1.1 (Plants), T3.1.2 (Pinned insects), T3.1.3 (Skins), T3.1.4 (Liquid samples), T3.1.5 (Microscopic slides), Task T3.2 (3D imaging), and T3.3 (Robotics). Each of these tasks or subtasks may produce test material, which can be made openly available in order to demonstrate the work of ICEDIG. Specifically, the following pilot projects have been foreseen in the Description of the Work for subtask T3.1.2 (Pinned insects):

    - Parts of multispectral imaging pilot study will be subcontracted for approximately 20,000 € to a research laboratory that owns the device "teraherz time-gated spectral imaging scanner". We will see if stacked labels in pinned insects can be read without removing them from the specimen. This could yield major cost savings in imaging. The project will process about 1,000 samples which will take approximately 1-2 months. Similar tests will be done at the Cardiff University using their available equipment.

    - Robotics pilot project will be subcontracted for approximately 20,000 € to a small company with experience in placing small cameras in robot hands, followed by taking large number of images of pinned insect specimen. These images will then be processed into a 3D model of the specimen. Labels underneath of the specimen should become readable when observed from different angles. This is new 3D imaging technology, and there are probably several robotics and image processing start-up companies that could do the job. One company will be selected using an open call for tender at the beginning of the project (MS16). It is estimated that the work takes 3-4 months.

- **The WP4 task T4.1 on automatic text capture may produce data on specimens.** This task will create a pilot project to test particular approaches. The task T4.2 may also test different strategies for data capture, which will produce data on specimens. This task will in addition coordinate the creation of a test dataset of herbarium specimens.

- **Transcription data in WP5, which belongs to task T5.2: Working with citizens to transcribe and enrich data.** This task will explore how to foster citizen participation and the emergence of new platforms beyond the existing ones. We will review existing platforms for volunteer transcription, such as Herbaria@home, DoeDat, DigiVol, Les Herbonautes, and Notes from Nature, as well as other platforms operating outside Europe, to evaluate which aspects of each system have been successful and which not. We will evaluate the quality of data produced from each system, including georeferencing, with the aim of identifying recommendations for improvement and further development (best practices). We will determine the motivations of participants, with the particular aim of increasing the diversity and number of participants. Furthermore, we will determine the need for internationalization of a citizen science platform to increase inclusiveness, but also to match citizens with the specimens from their own location/interest field/background. We will also look for novel usages of citizen participation to enrich data, for example, by extracting trait data, temporal distribution patterns, learning about field notebooks, the biographies of

collectors, and enriching gazetteers. To alleviate the coding effort for new websites, opening the source code of existing systems is a major help. Task 5.2 will review existing source code or code elements for transcription websites. To promote open access, it will build a comprehensive repository on a standard platform such as GitHub to allow public access on a stable version of the available sources codes and the associated documentation. Having a clear specification is the key to interoperability, and therefore the task will produce a specification to propose a simple way to publish transcribed data following existing standards for biodiversity data. It will also specify a lightweight interface for activity feed that can be used for a future "citizen transcription dashboard" on the DiSSCo website.

    - Specifically, a limited set of test data that may be transcribed during this work will be flagged as being the result of ICEDIG and will be made publicly available.

    - Despite the fact that code and documentation is not data *sensu stricto*, the platform will stick to this DMP regarding open access, security and reusability

- Citizens' digitisation pilot projects of WP5, which belongs to task T5.3: Digitisation of small collections. The biological collections of private collectors, amateur societies, and smaller museums and herbaria are numerous. Often they are very specialized and represent an important, but often unexplored or unknown resource. By bringing them together with other private and institutional collections they can contribute significantly to current data needs. As these collection owners have neither biodiversity informatics knowledge nor the resources to digitise and share their collections for science, the project ICEDIG will investigate solutions and procedures to incorporate these collections into the DiSSCo infrastructure. Pilot projects together with subcontracted citizen associations will be launched to test the ideas of how to best motivate and equip citizen collectors in digitisation.

    - Specific pilot exercises have been foreseen, which would be managed by local entomologist societies such as the Dutch Society of Lepidopterologists and the Estonian Lepidopterological Society. This is necessary to test procedures of training private collectors to digitise their own collections. The funds of 10,000 € will be used to hire a student to work as the trainer at the Society for about 4 months, and assist several amateur entomologists to digitise their collection. The data will be publicly shared through the GBIF portal.

    - In a related exercise, amateur entomologists which are users of the FinBIF portal and its Field Notebook service (see https://laji.fi/en/vihko), will be offered a service to print labels and unique identifiers for their own collection specimens. Currently, very few collectors are numbering or otherwise uniquely identifying and databasing the specimens in their private collections. Introducing such a practice would pave the way for digitising private collections before they enter public collections.

- WP6 will evaluate alternatives for a storage infrastructure at petabyte scale that may consist of combining different institutional, national, and European solutions. The WP will carry out tests and demonstrations of the data flows, storage systems, and access mechanisms in order to find the most viable solutions and combinations. These tests will be done in subtasks T6.3.1, T6.3.2, and T6.3.3 on national OSCs, EUDAT, and Zenodo, respectively. In each of these environments, digital

objects from the databases of project partners will be uploaded and the features of these storage environments will be tested and demonstrated. The description of the work is as follows:

- ***Subtask 6.3.1 National cloud infrastructures.*** In many European countries national solutions for open science clouds exist or are under development. The feasibility and role of these systems to store DiSSCo data will be assessed. Partners will gather information of the available systems and services from their countries and regions, which will be summarised in a document of services, capacities, and costs. In order to test these services one or more pilot projects in different countries will be carried out. The questions that will be clarified include data flows from digitisation facilities to these national level systems and further to European systems.

- ***Subtask 6.3.2 EUDAT infrastructure.*** In order to evaluate the infrastructure of EUDAT for DiSSCo, tests and demonstrations of the data flows, storage systems, and access mechanisms will be carried out. CINES will reuse its data service based on the existing Common Data Infrastructure developed by the EUDAT project. A storage capacity will be dedicated for the ICEDIG project as a B2SAFE (i.e., safe replication) service. The capacity provision during the ICEDIG project duration consists of maximum one hundred terabyte of disk and tape storage capacity accessible via the CINES's B2SAFE node in a similar way as for the EUDAT Data pilot Herbadrop. The Architecture consists of an ingestion service at CINES and optional replication centres (subcontracted by ICEDIG coordination to available EUDAT sites). All data ingested in the workflow include a validation stage to ensure that the data format is suitable for long term preservation. The workflow will follow OAIS recommendations and the repository will be DSA-WDS compliant to ensure that criteria for long term preservation are met. To demonstrate user functionalities, the task will build on the results of the Herbadrop data pilot. The Herbadrop pilot will be enhanced so that already processed data can be used to enhance the quality of the transcription (in cooperation with T4.2 and T5.3).

- ***Subtask 6.3.3 Zenodo infrastructure.*** In order to evaluate using the infrastructure of Zenodo for DiSSCo, tests and demonstrations of the data flows, storage systems, and access mechanisms will be carried out. Zenodo is based on exactly the same technology stack crafted by CERN to serve the Big Data needs of the High Energy Physics community. CERN uses this stack to power its own Open Data service (CODP), and as part of its mission to openly share the products of its research, it also used this stack to create Zenodo, within the OpenAIRE project, for all other researchers to use and store long tail science data such as over 175,000 published biodiversity images including links to external resources. In this task the storage will be extended for the large data needs of DiSSCo and connectors tuned for any domain specific needs. Zenodo is compliant with the open data requirements of Horizon 2020, the EU Research and Innovation funding programme and OpenAIRE. The data at Zenodo can be located via the ElasticSearch Engine. For each data set, a digital object identifier (DOI) is automatically assigned. Dashboards will be implemented to monitor input and content of the data locally. The pilot will include uploading of 100,000 images at a total of 10 TB uploaded in few big bursts, using the Zenodo API and then accessed from portals (WP4, WP5). This process will be monitored and documented.

Table 1. Summary of data types used by ICEDIG.

Work package and task	Objects	Number of Digital Objects	Volume (GB)
WP3 T3.1.2, T3.2 Pilots	Insect specimens in 2D and 3D	10,000	200
WP4 T4.1, T4.2 Data capture	Herbarium specimens	1,000	100
WP5 T5.2 Crowd-sourcing	Any	100	10
WP5 T5.3 Citizen digitisation	Insect specimens	10,000	200
WP6 T6.3.1 National OSC	Any	100,000	10,000
WP6 T6.3.2 EUDAT	Any	100,000	10,000
WP6 T6.3.3 Zenodo	Any	175,000	10,000
WP2, WP4, WP5, WP7	Questionnaires	1,000	0.01

Will you re-use any existing data and how?

Selected, existing images and data from the databases of the partner museums (UH, Naturalis, APM, UTARTU, NHM, MNHN, RBGK) will be used in specific tests, such as the storage tests in WP6.

The final kind of data that will be created is that which is information in project deliverables, which must be preserved, made accessible and passed on to subsequent persons working in DiSSCo.

What is the origin of the data?

These data have been digitised in diverse earlier projects.

What is the expected size of the data?

The size of the data handled by ICEDIG is quite small, such as less than 10 GB, except in the tests of the data infrastructure in WP6, where the project needs experience of managing large volumes of data, as explained above.

To whom might it be useful ('data utility')?

The data from these limited pilots will be useful for users and institutions who may be considering similar technologies in their digitisation and data management work. This applies in particular to the experiments carried out by WP6, but also the others. In particular, the digitised data from the experiments in WP3 will make apparent the quality of the digitisation results achieved with the new technologies. The data in the experiments of WP5 will be useful for the museums.

2. FAIR data

2. 1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism, e.g. persistent and unique identifiers such as Digital Object Identifiers (DOI)?

The data are discoverable through its DOI, if they have been published through the *Global Biodiversity Information Facility* (GBIF), or if they are deposited on Zenodo. However, there are data that are only available from institutional and national portals, which may not provide a DOI as of yet.

There is a recommendation by the CETAF ISTC of the form of a persistent unique identifier for specimens (see https://cetaf.org/cetaf-stable-identifiers).

What naming conventions do you follow?

Darwin Core (DwC) terms and *Access to Biological Collections Data* (ABCD) terms (see http://www.tdwg.org/).

Will search keywords be provided that optimize possibilities for re-use?

At Zenodo, keywords can be, and are routinely included in the metadata.

Do you provide clear version numbers?

In Zenodo, each deposit is clearly versioned, allowing adding multiple versions of the same digital object. Otherwise, no version numbers are provided.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Metadata are expressed in *Ecological Metadata Language* (EML), DwC, and ABCD.

EML is a *de facto* standard of the ecological informatics community, and supported by the *International Long-Term Ecological Research Network* (LTER, ILTER). It has been implemented in the *Knowledge Network for Biocomplexity* (KNB) and DataONE networks. It has also been implemented by the GBIF for all its resource metadata. EML can be used to describe datasets and projects. It does not cover the data itself.

DwC is a Biodiversity Information Standards (TDWG) standard, and can be characterised as a biological data extension of Dublin Core. DwC can be used not only for describing data resources, but also for "full data", i.e., location, time, observer, and species name.

ABCD also is a TDWG standard. It covers both resource metadata as EML, and full data as DwC. DwC and ABCD can be automatically cross-mapped.

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

All data produced by the experiments of WP3, WP4, WP5, and WP6, which has been described above, will be made openly available. This is, any imagery and results of automatic or computer-assisted human interpretation of the data, which can be seen in the imagery.

This does not mean that also the details of the equipment used and algorithms used in the interpretation will be made openly available, as these may contain proprietary information.

In Zenodo, the option exists to provide open access, embargoed access, closed access.

How will the data be made accessible (e.g. by deposition in a repository)?

The data will be deposited in the storage systems which will be tested by WP6, as appropriate (national OSC, EUDAT, Zenodo). Links from ICEDIG website will be provided to these storage systems.

By their service definition, the data stored at Zenodo remains permanently available. Permanent access to the data on national OSC and EUDAT tests is not foreseen. Data from the digitisation pilots may remain permanently available, if published on GBIF. These arrangements will be revisited after the data from the pilots has been created.

What methods or software tools are needed to access the data?

Web browser and/or *application programming interfaces* (API) offered by these storage systems, complemented by customized tools developed by users in specific domains. Zenodo provides basic robust, fast services. Anything on top of it is envisioned to be layered, and not necessarily part of the Zenodo infrastructure. For example, viewing and searching multiple images has to be handled outside Zenodo, e.g., by using https://ocellus.punkish.org/ that is currently being developed by Plazi for the domain specific Biodiversity Literature Repository.

Is documentation about the software needed to access the data included?

If accessed through the API, documentation will be needed.

Is it possible to include the relevant software (e.g. in open source code)?

Any such software has already been released by the providers of these storage systems.

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

The data will be deposited in the storage systems which will be tested by WP6, as appropriate (national OSC, EUDAT, Zenodo). Links from ICEDIG website will be provided to these storage systems.

Have you explored appropriate arrangements with the identified repository?

We have already explored the appropriate arrangements with the national cloud services in Finland (CSC), EUDAT through the work of Herbadrop pilot, and Zenodo through the work of the Biodiversity Literature Repository community.

If there are restrictions on use, how will access be provided?

There are no restrictions on use, except when CC BY-NC license has been chosen.

DiSSCO should address question of sensitive data (e.g. location of protected plants), but ICEDIG will avoid working with any sensitive data. If personal data is received in questionnaires, which ICEDIG will receive, such data shall be anonymised before making available outside the project.

Is there a need for a data access committee?

Because of the small scale of these experiments, there is no need for a data access committee.

Are there well described conditions for access (i.e. a machine readable license)?

The Creative Commons licenses supported by the GBIF will be used. These include CC0, CC BY, and CC BY-NC (see https://www.gbif.org/publishing-data). Zenodo supports a large array of widely used as well as domain specific, machine readable licences.

The owner of the data will determine which of these licenses will be used when data is posted on ICEDIG repositories. However, it is the project's recommendation to choose CC0 for data and CC BY for media, and avoid CC-BY-NC which has issues in some national jurisdictions.

How will the identity of the person accessing the data be ascertained?

Identity of the person accessing the data will not be directly ascertained. However, we expect users to follow the standard norms of scientific citation and use of the data in this context will be tracked through scientific citation.

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

The project will follow the standards and formats of the biodiversity informatics community.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

The metadata will follow the EML, DwC, and ABCD standards.

For images TIFF would be used for the originals, and JPEG for fast-loading derivatives on web portals.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

Standard vocabularies have been defined for some of the terms in above standards, but not all. The available vocabularies will be followed, e.g., ISO 8601 for dates and periods.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

The project will avoid generating its own ontologies and vocabularies.

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

The data will be licensed following the Creative Commons licenses CC0, CC BY, and by adding machine readable licences.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

There is no need to delay making the data available. Technical delays are possible, though, when making the arrangements with the repositories used.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

The data may be useful for scientists and digitisation projects, but as it is only created for testing and demonstration, the quantities and hence the usability will be limited. The tests of WP6 will use large datasets, but the data in those repositories (national OSC, EUDAT, Zenodo) are also available from other sources. If they are in multiple repositories, the respective alternative identifiers will be added to each of the deposit.

How long is it intended that the data remains re-usable?

Data digitised in the experiments of WP3, WP4 and WP5 will be migrated to GBIF at some point, and remain available in the long term. Also in the case of Zenodo, the data will remain available indefinitely.

Are data quality assurance processes described?

There is a specific task 3.4, which looks into the quality issues for image and another T4.2 which looks at data quality. The procedures created in these tasks will be tried in the data created in the pilot projects, where applicable.

Further to the FAIR principles, DMPs should also address:

3. Allocation of resources

What are the costs for making data FAIR in your project?

In case of the experiments of WP3, WP4, and WP5, following the standards will only result in savings, since we do not need to re-invent the formats. For the storage tests of WP6 there is a cost involved which depends on the size of the storage that will be allocated. There is a specific budget reserved for EUDAT (20,000€) and Zenodo (30,000€).

There may be at least one data paper and the costs for this will be covered by the project.

The experiences of the WP6 tests will be fed to WP8 which is looking at the costs of the DiSSCo research infrastructure. Although storage costs are rapidly coming down, and the physics community already talks of exabyte scale operations, the scale and cost of storing DiSSCo data at petabyte scale is beyond anything experienced by the scientific collections before. In ICEDIG, we need to understand accurately what those costs will be so that they can be planned for can plan in the Cost Book, which WP8 is constructing for DiSSCo.

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

This is included as service purchase in the grant agreement.

Who will be responsible for data management in your project?

The Coordinator University of Helsinki is responsible. There is also a specific work package (WP6) focussed on the data infrastructure. The creators of test data in WP3, WP4 and WP5 will manage their own data.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

In the case of Zenodo, this has been discussed. Long term preservation is actually in their business model. In the case of EUDAT, agreements of long term preservation have not been made because we are only carrying out temporary tests of data flows in and out of the EUDAT infrastructure. The same applies for the tests on national OSC, which tests, however, may evolve into operational systems. At national level either institutional (museum) storage will occur or it will be outsourced to some nationally agreed data centre. Control and dependency issues must be sorted out in this case.

The data from the pilot projects of WP3, WP4 and WP5 will be stored for long term in the institutional repositories of the partners.

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Institutional provisions are in place.

Is the data safely stored in certified repositories for long term preservation and curation?

This varies for each experiment, following the institutional provisions.

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

There is a specific task T7.1 “Identification, consolidation and harmonisation of national and European policy / legal frameworks”, and a related deliverable D7.1 “Policy component of ICEDIG project website”. There also is a specific WP10 for the ethics issues, as required by the European Commission during the grant preparation phase.

A provision is accepted that sensitive data (conservation status, governmental regulations, security issues) are not openly shared.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

The project’s questionnaires will not deal with any personal data, but if there is any, it will be removed before storing the data for any re-use. Where the data from project questionnaires will be stored has not yet been discussed. It first has to be determined whether it is necessary to keep such data after it has been written into deliverables.

WP10 might also address the issue of data on collectors of specimens. These data include dates and locations which people have visited and thus could be consider personal. However, much of such data is very old (even centuries) and can be considered historical and not personal. Where the line will be drawn needs to be considered in the DiSSCo DMP.

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

Each of the partners will follow their national and institutional procedures for data management, in addition to this ICEDIG DMP.