MEDIN Data Archive Centre Network – A Review of Future Funding Options

The MEDIN Executive Team, November 2010

1. Background and purpose

MEDIN is a collaborative and open partnership, established in April 2008, working to improve the management of marine data and information, and provide better access to the UK's marine data resources. Sponsors include government departments, research councils, environmental and conservation agencies, trading funds and commercial organisations. It operates under the auspices of the Marine Science Coordination Committee, and reports to that body.

The fundamental problem that MEDIN was established to tackle was that enormous amounts of data were being collected but in practice very little of this was available for reuse. There were over a hundred different holders of marine environmental data, with little or no coordination of standards and formats. This meant that discovering and accessing data was very difficult, and that even when sourced, the data were often unusable because of inconsistencies in standards and formats. Cowling (2005)[1] reviewed the situation and proposed a coordinated approach across the UK marine community. MEDIN established a three-pronged strategy to address this:

- Develop and maintain a common set of standards for data and metadata, essential to ensure that data can be discovered and re-used efficiently.
- Establishing a robust network of definitive, integrated, Data Archiving Centres (DACs) where data could be lodged for longer term curation and storage
- Providing a single central marine discovery portal through which all UK marine data sets can be searched and (eventually) accessed.

The DACS offer the capability to upload data, perform quality control and reformatting on ingestion (as necessary and agreed with suppliers) and then to provide secure long-term archival in a structured database with subsequent free access to the data owner and managed third party access to the data according to terms agreed with the data owner. In addition, through lodging data within the MEDIN DAC network the original data owners will be able to meet any metadata publication obligations from the INSPIRE initiative.[2] The DACS operating within MEDIN are subject to an accreditation process which ensures the DACs operate according to a set of best practice requirements.

It is important to distinguish between the capabilities of a Data Archive Centre and those of a database / dissemination tool. The MEDIN Data Archive Centres provide ***secure long-term storage*** for data, ensuring their availability to providers and the wider marine community for re-use in the future. They will also provide access to specific scientific expertise and advice on the data held within the DAC. A database is usually a local tool established to hold data for a specific project, often with time-limited support and limited (if any) inter-operability to other initiatives.

There are currently four accredited DACs within the initial MEDIN network: The British Oceanographic Data Centre (BODC) covering oceanographic components; the British Geological

[1] "Marine Data and Information – Where to now?" M. Cowling, Paper for IACMST, January 2005.
[2] The aim is also to ensure that INSPIRE data publication requirements will be met, but this cannot be guaranteed until the full requirements are published.

Survey (BGS) covering sea floor and sub seabed geophysical and geological data; the Data Archive for Seabed Species and Habitats (DASSH) dealing with the benthic data; and the UK Hydrographic Office (UKHO) holding bathymetric data. An expansion of coverage of the MEDIN DAC network is being considered to include meteorological, fisheries and heritage data. This will be achieved primarily by coordinating existing capability.

The outstanding issue with the DACs strategy is that it has never been formalised how the network would be funded in the longer term. This came to a head recently with regard to DASSH whose establishment was funded by DEFRA (from research spend) but once this came to an end there was no obvious viable long term mechanism to support it. As a temporary step, core MEDIN funds were used to extend the support for a further year while a longer-term solution was developed.

The purpose of this paper is to consider the situation around how the DACs are currently funded and to develop a proposal on a longer term, scalable, funding model.

2. Current Funding Arrangements

The following table lists the DACs that are currently accredited by MEDIN, with details of how they are currently funded.

DAC	Funding model	Comments and proportional funding	Funding Level Estimates for 2009-10
BODC	Established 1989 as NERC (Natural Environment Research Council) marine data archive. Funded through NERC normative budgets and programmatic support.	Core funding support to infrastructure and management - covers the basic IT (databases, web, ingestion and checking software), management, enquires and lab liaison activities. Data management agreements within projects and extra-budgetary resources from national and international programmes. May provide an active data management service to cruises etc.	(Full Economic Cost) Core costs: £1,200k (60%) Project data archiving costs: £710K (36%) MEDIN support costs £60K (3%)
DASSH	DASSH developed through R & D funding from Defra but no long- term financial support. DASSH also receives additional income through contract work from a variety of organisations (including CCW, MEDIN, Cefas, SEPA) that is complimentary to the overall work plan for DASSH.	Defra core funding covers up to 30 datasets a year. Additional project based funding has supported upload of further data sets . The remainder is an in-kind contribution from MBA staff	Core costs: £86.24K (75%) Project data archiving costs: £30K (20%) MEDIN support costs £15K (5%)
BGS	NERC normative budget and programmatic support via multiple funding routes. Multiple models exist in BGS and it is difficult to separate out the marine sector (e.g. what proportion of cost for a core store is the marine contribution) All income (from all sources) goes into the centre, and BGS data management is funded centrally from this budget.	Core Costs include archiving data collected by internal BGS projects, Costs have been supported from data sales and value added products.	(Full Economic Cost[3]) Core costs: £400K (75%) Project data archiving costs: £100K (20%) MEDIN support costs £24K (5%)
UKHO	The UKHO is a Trading Fund. The UKHO has agreed to provide nominal funding (cash and in-kind) for 3 years to set up and staff the Bathymetry DAC as part of the MEDIN network. Then the DAC becomes "business as usual" and will be staffed and run from within the UKHO Seabed (bathymetric) Data Centre. It is not envisaged that the additional MEDIN specific BAU staff and resource commitment will result in an increase in activity of more than 10% at UKHO.	The DAC functionality and remit have been specified so that almost all cost falls under core operations. It is anticipated that charges for archival may be made in the case of particularly large or complex data sets. Charges may be levied for data where media supply costs will be incurred	UKHO were unable to provide costs as the costs for the final delivery of the infrastructure are not yet available

[3] BGS costs estimate does not include costs for enquiry or delivery services, which are combined with non-marine data and are difficult to quantify.

The accreditation of two further DACs into the MEDIN DAC network is planned: a Met-Ocean DAC (the Met Office), and a Fisheries DAC (with three "Nodes" at CEFAS, Marine Scotland Science, and AFBI Northern Ireland). In each case core funding will be provided internally by the hosting organisation(s).

Summary

The total minimum annual running cost for three of the four existing MEDIN DACs is £2,625 k, £1,686 k for core capability, £840 k for project data archiving, and £99k for MEDIN coordination.

NERC provides core funding to BODC and BGS as National Capabilities to support marine and geological/geophysical science. Additional project based funding is provided to support the archival of new data sets within major projects.

The UKHO Bathymetry DAC meets an internal UKHO business need, but is also offered as a national capability as part of the MEDIN DAC network. UKHO has specified the scope and functionality of the Bathymetry DAC so that it can cover the core funding through internal resources.

The core costs for three of the current MEDIN DACS are supported through internal funding where an organisation has itself recognised the value of maintaining a data collection to its own business. This is also true for the two further DACs working for accreditation in 2010-11.

DASSH is the exception to this as it was established with Research and Development funding from DEFRA after a national requirement for a Marine Biodiversity DAC had been identified[4].

It is worth noting that for three of the four current DACS there is significant reliance on funding from projects (at least £840k in 2009-10), in which a project wishing to archive data pays for data management / archive costs specific to that data.

The third strand, the additional cost of acting as a MEDIN DAC: in meeting the MEDIN DAC requirements, reporting, generating and publishing metadata, attending and contributing to MEDIN meetings, has been estimated as a minimum of ~30-40 staff days per DAC per year, over £99k in total for the four DACs. Of course these activities represent added value to the DAC service, as they ensure data interoperability (lower costs in reusing the data) and the ability to meet metadata and data publishing requirements (for instance under INSPIRE).

3. Possible Models

There is a range of models that could be used to finance the DAC network. The table below lists the main options available along with the advantages/disadvantages associated with each. It should be stressed that the models are not necessarily mutually exclusive. Indeed, different aspects of the DAC functions could be funded by different mechanisms. We have identified four aspects to DAC activity:

- Core Capability: To support the core capability of a DAC, allowing it to carry out its routine functions. This includes: Infrastructure, administration, routine operation, responding to all

[4] see annex 3 for current situation with DASSH

user requests for data, an assumed "routine" volume of data archival per year. Includes meeting INSPIRE and UKLP metadata and (it is planned) data publishing requirements[5].

- MEDIN Coordination: Supports the additional effort required to act as a MEDIN DAC. This includes implementing DAC standards, attending DAC meetings, reporting to MEDIN.
- Additional archival costs: For archiving data beyond the "routine" volume included in core costs. Can include newly collected data, or data "rescued" from sources at risk.
- Retrieval Costs: To include the option that a charge could be levied for data retrieval. Clearly there can be associated costs if it is not possible to provide data automatically online (e.g. for high volume or complex data sets).

Model	Advantages	Disadvantages
1) Funded by the DACs, or the DAC parent bodies–who underwrite the function because they have a self interest in holding it.	Would allow for open-ended support to the MEDIN activity. Promotes data submission: attractive to data submitters in avoiding submission and management costs	May lead to gaps in coverage – model relies on finding a partner with a self- interest in holding and managing the data. Core funding rarely supports open-ended commitment to accepting new datasets – likely to require additional funding to support archival of new data (especially where greater levels of post-processing are needed).
2) Funded by data donors – data contributors would supply data to the DAC along with a financial 'dowry' to pay for the long term archiving..	Would be scale-related and support costs would be distributed. Clearly specified data management costs (whether through the DAC or to MEDIN). Does not rely on finding central resources to fund the network. Does not require an open-ended commitment from DACs. Forms part of the current model for NERC data centres and has proved sustainable.	May act as a barrier to data submission. May be an open-ended commitment for public bodies that they would be reluctant to support especially without effective audit. Fees may vary according to data type, some data will cost more to archive. Management overheads would need to be factored into the archiving costs, if core costs are not supported by the DAC host organisation.
3) Funded by data users – DAC takes a risk to hold the data and hopes to reclaim the costs by charging for access to the data.[6]	Distributes the costs to those using the data through licensing of potentially both raw and value-added data. Allows for varied licensing for different communities of users (commercial, research, academic, educational, charity). Good model for dealing with legacy data – i.e. The potential user pays for the mobilisation costs.	Charging levels would need to support the overheads and the management of legacy and non-requested data (e.g. dealing with data enquires). Will limit take-up and use of the data and detract from MEDIN proposed data distribution models. Disbursement of income to data generators rather becomes additional responsibility for DAC. Some data are free at the point of use and therefore would undermine the income generation for service provision. Users may choose to source data from elsewhere.
4) Subscription[6] – a variation on charging for use where potential users pay a	Attractive to the data users subject to the scale of the charges levied, as it makes access simple with a one-off or periodic subscription. Potential simplification of portals subscription	May not generate sufficient funds to operate across all DACs. Becomes more limiting as the number of DACs increases if users have to subscribe to multiple

[5] An absolute commitment cannot be given until the details of the regulation are known.
[6] Options (3) and (4) are no longer viable for the vast majority of the data held in MEDIN DACs because of recent developments, including the Government's transparency agenda and open government license. Also NERC has adopted a policy not to charge for access to NERC data.

flat rate per annum.	if managed through a single MEDIN portal. Potentially simplifies the financial planning for the DAC.	portals. Limits public access to the data which may be a problem for some data generators. Though note that an obligation to make data accessible does not preclude charging for the data. Users may choose to source data from elsewhere.
5) Funded by MEDIN – DACs are funded through a central fund administered by MEDIN.	Would provide longer term security and certainty of support if MEDIN is seen as a stable structure – with predictable resources and support to infrastructure. Allows MEDIN to set the protocols and distribution mechanisms for marine data.	MEDIN is a 5-year funded programme and therefore may not be as stable as existing DAC funding programmes. Adopting MEDIN support as sole-funding may weaken sustainability of the DAC for its other (non-MEDIN) activities/ international activities etc. Would need to be established with appropriate service level agreements with the DACs. Undermines the current arrangements already established for internal business reasons (e.g. by NERC).
6) Programmatic support.	Allows flexible development of services that are tailored to user needs.	No assurance that the programme funding will be secured or that the topics can meet the service level needs of MEDIN. Funder led initiatives may limit the scope of the access and DAC control of development. Unlikely to provide a stable long term mechanism
7) Funded from central government	Allows for a consistent approach across the network (assuming such an approach can be agreed). Provides a mechanism to ensure long -term support.	Funding may be vulnerable to political climate. Current financial environment means that central funds are unlikely to be available. Cost burden is not shared by all who stand to benefit.

In summary all of the models have their own risks. Funding through an ingestion cost route risks the data not being published as the provider may not be willing to pay this cost (particularly if they have no obligation to do so). Deriving the costs from charging for the use of the data (either directly or through a subscription) potentially reduces the amount of use that is made of the data and hence partially undermines the whole business case for MEDIN. However, expecting a DAC to accept all data, and to service all data requests without any charging is effectively that DAC asking for a bottomless commitment. Making an arguable case for central funding to underwrite a national capability is difficult because of the large cost involved, and as it would involve altering existing arrangements (for BODC and BGS) that have proved durable and sustainable.

The need for long term security and sustainability is also a very important factor to bear in mind. This is important both from the perspective of MEDIN itself but also from the perspective of the data providers who want to be confident that time and effort expended in getting data into the DAC network secures data for the long term. Learning from the experience around DASSH the investment in 'pump-priming' such a facility should really only be committed when there is a clear and agreed longer term mechanism to fund the facility.

Currently, the major contributors to the ***visible*** costs of data archiving for Marine Data include:

a) The Natural Environment Research Council which pays through its core funding of BODC and BGS, and the top-slicing of research projects to support data-archiving costs.
b) UKHO, which is underwriting the costs of operating the UKHO Bathymetry DAC
c) DEFRA and the Scottish Government which have underwritten the large part of establishing the Marine Species and Habitats DAC at DASSH.
d) MEDIN sponsors who are supporting the costs of establishing common standards across DACS.

It is expected that in the future these visible costs will be sustained at a similar level, but that the results of improved coordination will be significant reductions in less visible, indirect costs, which are borne by the whole community, such as

- The costs of data management and publication on organisations that do not archive their data within DAC network.
- Meeting the INSPIRE / UKLP requirements for publishing metadata and data publication.
- The additional costs on third party organisations wanting to access data from outside the MEDIN DACs, which include:
    - Identifying all potential sources of data and identifying which data are of interest, described in non-standard metadata.
    - Combining data from different sources, often provided in different non-standard formats, with different levels of QC and inconsistent geo-referencing.
- The additional costs in re-surveying for data that are not known about.

4. Recommendations

Overview

MEDIN aims to act on behalf of the whole marine community, covering a wide range of interests and organisations including:

- **Research** – obviously one of the primary consumers of data relating to the environment is the research community. This is supported by the fact that the Natural Environment Research Council (NERC) already invests quite heavily in the archival of data through BODC, BADC and BGS.
- **Natural environment conservation and planning/management** – this is one of the primary drivers behind the establishment of MEDIN (originally identified through the first State of the Seas report). Given that this area is a devolved function the lead is shared between DEFRA, Scottish Government and Welsh Assembly Government. At present there is no direct long-term support to the archiving function in this area other than an expectation that those bodies that are funded through Grant in Aid (GIA) will manage the data that they are responsible for.
- **Transport** – another significant driver relates to transport and more specifically the need to have a maintained picture of the physical features of the marine environment (both natural and man-made).
- **Conservation of the historic environment** – while there are benefits to this community, overall the benefits are not as strong as those listed above.
- **Offshore Operators and Developers** – The commercial sector is a key source of data, but tends to fund its own data collection and management. This sector includes the oil and gas industry, the aggregates industry, and developers / operators of renewable energy installations

There is not a single central source of funding of sufficient size to meet the data archiving requirements of all these sectors of the community. In any case some arrangements are already in place to meet specific sector requirements, which are reasonably secure. Therefore, rather than revise these arrangements, it is preferable to build upon them in a way that allows a consideration of the wider national need and ensures an equitable contribution from those sectors who stand to benefit the most.

Proposal - A Collaborative Funding Model

A co-ordinated UK wide funding approach is proposed to ensure that a core operational Data Archive Centre (DAC) capability exists to meet key national needs. The sustainability of this approach requires an acceptance that data archiving costs money and that those funding data collection include in the project costs the funding necessary to support the archival of the data, in return for

the services provided by the Data Archiving Centres. MEDIN acts to set the service requirements and standards and to monitor the performance of the DACs in meeting these requirements.

There are four aspects to the recommended approach:

- **Core DAC Capability**
  We propose a collaborative approach to provide support for the necessary "core" national DAC capability, which includes infrastructure costs and some routine data archiving. Thus it would be expected that core DAC funding is provided by organisations with a strategic interest in the existence of a national DAC capability for specific data types. It is recommended that a DAC funders group is formed to provide high-level oversight to ensure core DAC capability meets national needs.

- **MEDIN coordination**
  MEDIN acts to ensure common standards and service provision across the MEDIN DAC network. It is recommended that the cost of MEDIN coordination activities is shared between MEDIN Sponsorship funds and the DACs themselves. The sponsors have signed up to the MEDIN principle that marine data should be more easily accessible and re-useable, which is the objective of this strand of DAC activity.

- **Additional Archive Costs**
  It is proposed that the costs of archiving newly collected data would be funded by the data providers themselves. Thus data providers pay one-off fees to the DACs in return for the services provided, which include: data quality assurance and reformatting, data upload and storage, and publication of the data to third party users (this includes meeting obligations to INSPIRE and UKLP). It is recommended that these costs are agreed with DACS and costed into data collection projects from the beginning.

- **Data retrieval**
  The MEDIN DACs will provide data access the original data provider at no cost. MEDIN DACs will manage third party access to data sets according to terms agreed with the data provider. If no constraints are required, data will be made available to third parties at no cost, beyond any necessary to cover costs of retrieval / provision.

The logic behind this recommendation is that:

- Organisations with a strategic interest in maintaining a national marine DAC capability are involved in the collaborative funding of this capability.
- Organisations who have signed up to the MEDIN principles of improving management of and access to marine environmental data support the coordination role provided by MEDIN.
- Those who fund the collection of marine data have a responsibility to ensure the data are available for re-use.
- The best way to ensure data are most widely re-used is to provide free of cost access where-ever possible.

More detail, including anticipated costs, is provided below:

CORE DAC funding

The core DAC capability is defined as that capability necessary to allow the DAC to carry out its routine functions. This includes: Infrastructure, administration, routine operation, responding to all user requests for data, an assumed "routine" volume of data archival per year. Includes meeting INSPIRE and UKLP metadata and (it is planned) data publishing requirements[7].

A coordinated funding approach to ensure core operational DAC capability exists to meet key national needs is proposed. It is recommended that the current funding arrangements stay in place for BODC, BGS and UKHO, as they are secure and sustainable in that the DAC function is required to meet particular organisational business needs. The situation for DASSH is detailed in Annex 3, but it is anticipated that long-term core funding will be provided by a consortium of organisations with a strategic interest in supporting a DAC for marine species and habitats.

The key recommendation is that a DAC funders group is formed, which includes (but is not necessarily limited to) key government departments who have a relevant policy requirement for marine DAC capability. This high level group would meet to review and respond to any funding risks to the existing DAC capability, and to respond to any recommendations to provide additional coverage of the DAC network. By taking a high level overview, existing funding commitments (e.g. by NERC, UKHO, DEFRA, Scottish Government) can be taken into account and the financial burden of any additional necessary capacity shared equitably.

The associated costs of supporting core DAC Capability are of the order of £100k- £1M per year per DAC, (ranging from £87k for DASSH to £1.2M for BODC).

The funders group would meet approximately annually and review:

- The principles on which the co-ordination is based. In addition to the provision of the core DAC functions (secure long-term storage for marine data, the capability to upload and retrieve data, to make available clear searchable information on their data holdings and to provide a source of expertise for the management of marine data), these would include:
    - Data to be made available to users at no additional cost and can be freely re-used except where the actual data owner has imposed restrictions.
    - Data users will be asked to acknowledge the source of the data in all uses.
    - Discovery metadata will be published to the MEDIN marine discovery portal, and to other portals as directed by the data provider. This will satisfy any metadata publication obligations the provider may have under INSPIRE and data.gov.uk.
    - Costs for archiving data sets should be agreed in advance between the data provider and the DAC. These costs will cover necessary Quality Assurance, cleaning and reformatting, and upload to a structured secure data base.
- Any gaps in the network and the business case for why these gaps should be addressed. This would be a contribution from the central MEDIN secretariat.

[7] An absolute commitment cannot be given until the details of the regulation are known.

- The level of funding that each department/partner is investing in the overall archiving function and whether the relative contributions of each partner should be rebalanced and potentially certain DAC capabilities ceased (e.g. where little or no use is being, or is likely to be, made of the data but substantial costs are being incurred) or new ones established. This would have to bear in mind any obligations to publish data under INSPIRE and data.gov.uk

- How well the network is performing. The group would review a short standard report provided by the MEDIN DAC Working Group covering:
    - The volume of data that has been archived in each DAC and the costs associated with this.
    - The level of use that is being made of the data archived (this would be both in terms of academic publication but also integration into other planning, management and reporting functions undertaken across the UK and potentially beyond the UK). Again this would rely on reporting from each DAC.
    - Any issues or needs that have arisen either from the perspective of the DAC or feedback or comments from data providers or users (either submitted to the DAC or MEDIN itself).

There would also be a need to establish an agreed process for considering proposals to expand the coverage of the MEDIN DAC network. This would include:

- Analysis of the specific gap in coverage, agreement that this represents a significant problem and must be addressed.

- Review of options to build on existing capability or for an existing DAC to widen its remit. Only in exceptional circumstances would a completely new capability be considered. Consider expertise, capability, service offered, costs and requirements for funding.

- If none of these are satisfactory, consider if there is a case for a new DAC. Should not be initiated unless a long-term funding steam can be established.

Note that one of the current requirements to be met by a DAC to achieve MEDIN accreditation is to provide a long-term stewardship plan, which includes a statement on how the DAC is financed and for how long, and the actions that would be taken in the event that DAC becomes unsustainable.

Note also that in 2010 MEDIN is planning an expansion of the MEDIN DAC network to include a DAC for marine meteorology (based at the Met Office, with core funding provided by the Met Office), and a DAC for fisheries data (with nodes at CEFAS, Marine Scotland, and AFBI). Possible arrangements for heritage data are being discussed.

MEDIN Coordination

MEDIN Coordination is required to ensure that DACs within the MEDIN network meet the service requirements and the agreed standards so that those archiving data within the DAC network can be assured of the service being provided.

The required capabilities of DACs within the MEDIN framework are:

- To ensure the secure, long term, curation of key marine data sets, according to best practice and to relevant national and international standards.
- To make available clear, searchable information on their data holdings, by the generation and publication of metadata on the MEDIN portal.
- To form the first point of call of expertise for the management of marine data.

In autumn 2010 MEDIN will begin to publish metadata to the data.gov.uk portal, satisfying UKLP and INSPIRE metadata publishing requirements. When the INSPIRE data publishing requirements are finalised MEDIN will investigate how these requirements could be met through the MEDIN DAC network.

The detailed requirements for accreditation as a MEDIN DAC are given in Annex 1. It can be seen that these conditions are strenuous and place significant obligations on the DAC. As a further condition of its accreditation, each MEDIN Data Archive Centre is required to provide a short annual report so that sponsors can assess how well the DAC framework is operating, and to participate actively in the MEDIN DAC Working Group to support further planning and coordination.

The costs of the MEDIN DAC work stream for 2010-11 are £180k, supported by In-kind effort from DACS estimated at a value of £99k (approximately 30-40 days staff effort per DAC).

The proposal for future operation of the MEDIN DAC work stream is that the DACS themselves take a more leading role, rotating the chairmanship of the working group between them and that the contribution of the MEDIN core team reduces to a coordination, planning and secretariat role. It is proposed that the contribution of staff effort by the DACs to this work stream is directly supported from MEDIN funds to the equivalent of 20 days per DAC per year, with the balance being provided by DAC in-kind support. The cost to MEDIN funds would be ~£60k per year, and has been included in the projected annual budget for the DAC work stream of £132k per year.

One-off Data Archiving Costs

The third aspect of the recommendation is that all data providers will pay one-off data archival costs to archive their data within the MEDIN DAC network. This includes newly collected data, or data "rescued" from sources at risk. The only exemptions are for those data (usually collected by the DAC funding body) whose routine archival is included in the core costs.

For new data collection projects outside the core activity there would be a one off archiving costs retrieved by top-slicing the project costs (akin to the NERC model).

"Data Rescue" projects would have to be funded, so proposals would have to be put together and presented to funding bodies.

Perhaps the most important factor in ensuring future sustainability of the MEDIN DAC network will be gaining the acceptance by data providers that data archiving costs money, and should be factored

in to projects involving data collection from the very beginning (this is a fundamental tenet of the MEDIN "Data Clause" – annex 2).

Costs will vary depending on data type, and there is potential scope to reduce costs by using standard formats. For 2009-10 these costs were estimated to be at least £840k across the MEDIN DAC network.

Data Retrieval Costs:

Provided that the original data owner has placed no access constraints, it is recommended that the MEDIN DACs adopt the basic principle is that ***all data will be made available at no cost,*** beyond that necessary to cover costs of retrieval / provision, and that no constraint is placed on subsequent use. This is in line with the recently published "Open Government Licence" and NERC data policy.[8] In the case of online data provision these costs would be zero, costs would only usually be charged in the case of high volume complex data sets requiring manual intervention and the supply on hard media (CDs, DVDs, Hard Drives).

5. Key tests

In this section we assess how the recommendations stand up to key tests of Data Security, Cost Effectiveness, Sustainability and Extendibility.

Data security

- Access control: Control over access to data is ensured by the DAC standards. One of the MEDIN requirements (Annex 1) is for DACS to manage access control according to conditions established by the data owner. Thus where no constraints are required, data can be made freely available, alternatively if access to data is to be restricted to specified list of users, DACs can also enforce this. This type of functionality is a core requirement.
- Security against accidental or deliberate deletion, etc. Again one of the MEDIN DAC requirements is to have systems in place to ensure security against deliberate, accidental deletion or other events such as fire. All DACS are required to have secure back up processes and disaster planning in place.

Cost effectiveness

- The overall visible costs for maintaining the DAC network are projected to remain roughly the same, as the current core funding arrangements for DACS are not changed. The MEDIN DAC work stream costs will reduce by approximately £50k from 2010-11 to 2011-12.
- It is expected that internal data retrieval costs within MEDIN partners will reduce as it becomes easier to source and use data from third parties that are accessed through the MEDIN DAC network.

[8] http://www.nationalarchives.gov.uk/doc/open-government-licence/open-government-licence.htm

- In addition to ensuring the more efficient access to and reuse of data, the overall cost of marine monitoring should be reduced, as the number of unnecessary re-surveys should be reduced.

Sustainability

The recommended approach offers long-term sustainability in the following ways:

- The DAC core function for specific data types is supported by those with a long-term strategic interest. (e.g. NERC for BODC and BGS; UKHO for Bathymetry; DEFRA and the Scottish Govt. for marine biodiversity)
- The cost of archiving new data is included in the initial calculation of costs for new projects involving the collection of marine data.
- No organisation is being asked to make open-ended commitments to archive and disseminate data with no cost related funding support.
- Extension of coverage of the DAC network will only be considered if long-term funding is assured (a requirement for DAC accreditation – Annex 1).

Extendibility

- The DAC core funders group is established to consider the balance of national interest versus cost.
- Any proposals to provide new core capability only accepted where there is long-term funding commitment from an organisation, or group of organisations, with strategic interest.
- Extension of coverage of existing DACs will be the default first preference, or to build on existing capability.

Equitable

- The cost burden for each aspect of the DAC network falls on those who have a strategic interest or are being provided with a service.
- Core function is supported by organisations with a strategic interest in a specific type of data.
- MEDIN coordination costs are supported by all MEDIN sponsors – who have signed up to the MEDIN objectives of improving management of and access to marine data.
- Costs of data archival are supported by the data providers, who are encouraged (through the Data Clause) to include provision for this cost in the initial project planning phase.
- DACS to respect and manage access according to any conditions on data access requested by the original data provider. Otherwise data to be accessible to all from the DAC network at no cost above any cost of recovery or media cost.

Annex 1: Marine Environmental Data and Information Network (MEDIN): Accreditation Process for Data Archiving Centres

Introduction

A key objective of MEDIN is to establish an operational network of linked marine data archive centres (DACS) to provide secure long-term storage for marine data. This network will provide the capability to upload and retrieve data. Data contributors should have free access to their data managed within the DAC framework.

The required capabilities of DACs within the MEDIN framework are:

- To ensure the secure, long term, curation of key marine data sets, according to best practice and to relevant national and international standards.
- To make available clear, searchable information on their data holdings, by the generation and publication of metadata on the MEDIN portal.
- To form the first point of call of expertise for the management of marine data.

MEDIN has established an accreditation procedure to govern the process by which new Data Archive Centres are included into the network. Once accredited DACs must provide annual reports for the MEDIN Sponsors.

Accreditation Process

There are six stages to the accreditation process, finishing with formal approval by the MEDIN Executive Team:

- *Initiation / Preparation*
- *Response to MEDIN DAC Requirements*
- *Review of DAC Response*
- *Updated Response to MEDIN DAC Requirements*
- *Recommendation from Expert Panel*
- *Accreditation by MEDIN Executive Team*

The first, preparation stage can take up to several years. Subsequent stages should take between 8-12 weeks before the final accreditation by the Executive Team.

Once accredited, the status and performance of DACs will be reviewed annually as part of the annual review process.

The expert panel who review the DAC response and provide the recommendations to the Executive team will only include members who are independent of the DAC being considered.

Initiation / Preparation

Involvement: MEDIN DAC working group, experts and organisation proposing to host a DAC.

Description: The MEDIN DAC Working Group identifies the need for a further Data Archive Centre within the MEDIN network. Working with interested parties the DAC Working Group proposes an outline scope, remit and *modus operandi* for the new DAC.

Duration: This part of the process can take between 6 months and 2 years, as it requires a consensus to be established between interested parties, and perhaps business plans to be developed.

Response to MEDIN DAC Requirements

Involvement: DAC host organisation

Description: The new DAC provides a detailed response to the list of MEDIN DAC requirements as detailed in the Appendix.

Duration: 2-4 Weeks

Review of DAC Response

Involvement: DAC Expert Panel

Description: An expert panel appointed by the DAC Working Group / DAC Executive Team reviews the DAC response and identifies where (1) further information is required, and (2) where the proposed arrangements do not meet MEDIN requirements.

Duration: 2 Weeks

Updated Response to MEDIN DAC Requirements

Involvement: DAC host organisation

Description: The DAC responds to the reviewers' comments and updates its arrangements as necessary (or proposes a work programme to do so).

Duration: 2-4 Weeks

Recommendation from Expert Panel

Involvement: DAC Expert Panel

Description: The Expert Panel provides recommendations to the MEDIN Executive Team on the DAC application

Duration: 2 Weeks

Accreditation by MEDIN Executive Team

Involvement: MEDIN Executive Team

Description: The Executive Team considers the Expert Group's recommendations and:

(a) Confirms accreditation of the DAC.

(b) Confirms accreditation of the DAC but recommends specific actions to be taken by the DAC to meet requirements.
(c) Postpones accreditation of the DAC until specific actions are taken.
(d) Recommend that an alternative solution be found to provide Data Archiving facilities for the data categories under consideration.

Duration: At MEDIN Executive Team Quarterly Meeting

Marine Environmental Data and Information Network : Requirements for Data Archiving Centres

This document lists the requirements for an organisation to become a Data Archive Centre (DAC) under the Marine Environmental Data & Information Network (MEDIN). It also provides further explanatory information for each of these requirements to ensure that potential DACs are clear as to the evidence needed to be provided in order to be accredited as a DAC.

Requirement	**To be accredited DACs must provide**
ORGANISATIONAL FRAMEWORK	
Generally exhibiting evidence of expertise and a track record in the scientific area of the data	DACs should describe the range and length of expertise of both the organisation and their staff. In addition, details of data sets or products available can also be provided Any appropriate affiliations (e.g. national or international bodies, etc.) should also be noted.
Committed to provide sufficient resources for defined period of time and plans for transition if and when it ends	In order to be accredited, a DAC must provide evidence that it is hosted by a recognised institution (ensuring long-term stability and sustainability) and that it has sufficient funding, including staff resources, IT resources and a budget for attending meetings, ideally for a 3 to 5 year period, and this information should be updated regularly.
Committed to return of data holdings to originators, or lodging with an alternative and suitable repository, if the DAC becomes unsustainable	A long-term stewardship plan should be available including: • A statement on how the DAC is financed and for how long. • Action that will be taken in the event that the DAC becomes unsustainable
Provide annual report as specified by MEDIN	Accredited DACs should provide an annual report to MEDIN according to the pro forma provided by MEDIN. The report comprises 4 sections as follows: • A short summary of the remit and status of the DAC • An overview of activities and developments in reporting year • Key Targets for the next reporting year • Report any changes against the specific MEDIN DAC requirements (in particular referring to any requirements placed as a condition of accreditation Other suggestions for future reports might include: • Key Performance Indicators • Statement on readiness for INSPIRE compliance

QUALITY CONTROL AND MAINTENANCE	
Adherence to MEDIN Discovery Metadata Standard and appropriate international principles	MEDIN DACs need to provide evidence of adherence to these principles. Further information and links are given below. The MEDIN Metadata Discovery Standard must be used to record details of data sets. The fields used in the standard are compliant with other international conventions (INSPIRE, ISO19115), which means that the details can be transferred easily between organisations and queried by the MEDIN portal. The Metadata Discovery Standard also conforms to the GEMINI2 profile. Publication of metadata in the MEDIN Metadata Discovery Standard and made available to the MEDIN Discovery Portal meets both INSPIRE compliance and UK Location Programme requirements for discovery services. ISO 19115 (Geographic Information - Metadata) is an international standard that sets out a number of metadata fields for describing spatial information datasets. ISO 19139 (Geographic Information - Metadata - XML schema implementation) is the standard that aims to define an XML encoding for the metadata elements defined in ISO 19115. The UK GEMINI Discovery Metadata Standard is a defined element set for describing geo-spatial, discovery-level metadata within the United Kingdom. It is derived from, and therefore compliant with, ISO 19115 Geographic Information – Metadata and the UK eGovernment Metadata Standard (eGMS). GEMINI was originated by the Association for Geographic Information and is currently being revised to produce GEMINI 2. A number of tools and documents to assist in creating MEDIN-compliant metadata are available from the Standards section of the MEDIN web-site.
Data collection according to defined quality principles and accepted procedures	MEDIN DACs need to provide evidence of defined quality principles and procedures. DACs may also be able to advise on data collection procedures and should be able to direct data collecting organisations to appropriate standards, where these exist. MEDIN is also in the process of deriving data guidelines comprising requirements as to what must be recorded when data of a certain theme is being collected. This allows easier reuse of the data in the future. For example, if benthic invertebrate samples are collected, the instrument used to sample, the sieve size and taxonomic list used to record species should also be stated and use common lists of terms. MEDIN approved data guidelines are available from the standards pages of the MEDIN web site. Where MEDIN data guidelines do not already exist, it is recommended that the resources available on the other marine data standards web pages should be used. Provision of advice and feedback to the original data collectors is valuable, covering information to be recorded alongside data, established quality assurance procedures to be used, etc.
Quality assurance of the collected data	MEDIN DACs should provide summaries of any quality assurance processes and algorithms that are in place. This should not be a detailed description of how the algorithms work but a broad summary of the checks that are run and, for example, whether data are visually inspected. The summary should include details of how any

	issues are resolved (e.g. are they returned to the data provider for rectification, fixed by the DAC, noted by quality flags in the data file and/or included in the accompanying metadata). In addition, details of any Quality Management System (QMS) or accreditation schemes implemented by the DAC should be provided. Where data have been collected in line with nationally or internationally agreed standards this should be indicated. For example: • Quality Assurance of Information for Marine Environmental Monitoring in Europe (QUASIMEME) • Biological Effects Quality Assurance in Monitoring Programmes (BEQUALM) • National Marine Biological Analytical Quality Control Scheme (NMBAQC) • ISO9000 accreditation • Data collected to internationally agreed standards within major scientific projects (e.g. JGOFS protocols and standards) Where guidelines and standards are in use these should be mentioned. For example, the ICES Working Group on Marine Data Management has developed a series of "Data Type" guidelines, which have been designed to describe the elements of data and metadata important to the ocean research community. These guidelines are targeted toward physical-chemical-biological data types collected on oceanographic research vessel cruises.

Committed to advising third party organisations collecting similar types of data on procedures, and providing data-banking (warehousing) and curation facilities for such similar data from other sources	Short description of DAC
	• Short description of the remit of the DAC including the data types held and those accepted from external parties for archiving.
	• Licensing terms
	• Standard agreements covering:
	  • Transfer of a copy of data to a DAC
	  • Transfer of ownership to DAC
	  • Use of the data held by DAC by external users
	Format requirements
	• Note that these are aspirational for new data being collected which needs to be submitted to a DAC. It is not intended that all historical data would need to be converted to these formats before acceptance by the DAC. Historical data needs to be addressed on a case by case basis.
	• At least one, but potentially more, format(s) that data can be submitted to the DAC.
	• Details of the process for establishing or agreeing alternative formats.
	• The format description would need to cover both format and syntax.
	It may be advantageous for the provider to submit data in their own format provided this is properly documented perhaps along with some sort of index of the data.

TECHNICAL INFRASTRUCTURE	
Databasing and banking with appropriate metadata standards	MEDIN DACs should provide documentation of their working practice and procedures. This should include: Information on the technical metadata for all holdings. • Descriptions of the data structures (both entities and attributes) within which the data are stored

	• Explanations of any lookups not obvious from the data holdings directly • Locations of data holdings on the network or other physical locations • Information on metadata schemes • Editorial advice on the content expected in each mandatory field of ISO xxx • List of any topic specific additional fields and accompanying editorial guidance • Information on georeferencing standards in use
Auditable process for long term custodianship and updating of data sets, with appropriate disaster planning	MEDIN DACs should have a security policy describing how the data holdings are protected from both malicious and accidental loss. Note that the security policy should exist but should not be made public as it potentially exposes vulnerabilities. A policy should include the following: • How the holdings are physically protected (e.g. how access to the building is controlled, how secure the building is, who has access) • Access to the network (if the holdings are accessible from the network) – what is the access policy, how is user access limited and by who, whether there is an internet link and details of how the firewall is configured and altered, how machines are patched, which users can log on to particular machines, policy on passwords (e.g. how often they are changed and how secure they need to be) • Policy when staff leave organisation • Database policy – how users are established, what rights they have, how often administrator passwords are changed, what control is there over allowable passwords How the data holdings are backed up – how often, where are the backups stored and how long for, how protected are the backups (e.g. fire proof safe, stored securely off site, who has access)
USER ACCESS AND COMMUNICATION	

Committed to, and focus on, customer service	DACs should provide information on:
	• Response times to enquiries for data and information
	  • Description of aimed service level for responding to user requests (where these are cannot be met on-line).
	  • Whether an Enquiries or Help Desk is available
	Details of surveys of customer satisfaction undertaken

Committed to raising awareness of the holdings and promoting the use of the data	Describe facilities available at the DAC to discovery data holdings: • Details of how the data can be searched or interrogated by interested users (e.g. On-line metadata search, physical access on site etc) • Short summary of any on-line search functionality Describe other search facilities used, e.g. • Discovery metadata available through the GI Gateway, National Biodiversity Network, UK MED Directory/EDMED, etc. The DAC should provide an indication of participation in conferences and exhibitions; production of promotional leaflets, flyers and articles In addition to the activities above the DAC should provide information on: • Data products available

	• Linkages with other organisations who use the data for generation of products • Current projects aiming to increase and promote data use • Statistics/metrics indicating data usage
Making datasets freely available wherever possible (not necessarily at zero cost)	MEDIN DACs should have a policy on data access. In general DACs should aim to make data sets freely available, although it is recognised there may be restrictions on access to data for a number of reasons including national security, commercial confidentiality, for scientific research to allow the principle investigators and their co-workers to exploit the data in the first instance. However, release of data to the wider community after a period of 1-3 years from data collection should be strongly encouraged. Metadata should be made available at zero cost and data should be made available at zero cost where ever possible. The data access policy should include the following: • Details of what can / cannot be obtained on-line (e.g. metadata only, full dataset download) • Licensing arrangements • The format(s) that data can be provided in • The media used for providing data (if data are not on-line) • Costs associated with data provision (or cost scales) – including cost of media as well as staff time Wherever possible, data policies should be in accordance with internationally agreed data policies (e.g. IOC Oceanographic Data Exchange Policy, GOOS Data Policy, WMO Resolution 40, ICES Data Access Policy, etc.)

Assumptions

1. It is accepted that there may be instances where there is more than one copy of a dataset within the MEDIN structures but that there will be one MASTER (original) version, held by the originator or transferred to a DAC
2. It is accepted that there may be instances where datasets of similar type are held in separate DACs
3. It is accepted that there will be a range of different levels of value added and commercial activity with the MEDIN DACs
4. There are Funders of Data Collection, Contributors of Data, Holders of Data and Users of Data in MEDIN (all subject to relevant sets of requirements) as well as DACs; these roles are not mutually exclusive.

Annex 2: Marine Environmental Data and Information Network (MEDIN) - Proposals for common contract clause for data collection

Introduction

The Marine Environmental Data and Information Network (MEDIN) aims to promote best practice in data gathering to ensure that data are properly archived. To ensure that (public sector) research and survey commissioning bodies (Clients) adopt this best practice and have a contractual basis for the data gathering programs they commission from Contractors (or Tenderers), MEDIN has developed a style of standard clauses that can be used in tender specifications, so forming a fundamental part of the contract from the start. This will ensure that data management best practice and its associated costs are addressed by Contractors (tenderers) at the tender compilation stage.

This document collates the experience from the use of existing "data clauses" in contracts across the marine community in the UK with a view to providing a standard clause with guidance for implementation.

Requirement

Ideally, data collection contracts should ensure that the following issues are addressed:

- The application of, and documentation of, appropriate standards during data collection.
- The generation and provision of metadata in an agreed standard format.
- That provision is made for the secure long-term archival of the data.
- That ownership, Intellectual Property Rights, and terms and conditions for third party use of the data are clearly and unambiguously established and documented.

Experience so far

There have been two main uses of a data clause to date:

1. DTI/BERR adopted (through their main contractors, Royal Haskoning) a Marine Data Acquisition clause in a contract for their SEA surveying work in the N Sea.
2. The Channel Coastal Observatory (CCO) applied the approach detailed at: http://www.channelcoast.org/data_management/online_data_catalogue

MEDIN has discussed the effectiveness of these approaches with Royal Haskoning and the CCO. Royal Haskoning advised that they were comfortable with the application of the clause, as it represented best practice. The CCO confirm that establishing the requirements at the tender stage ensures that data management issues are addressed properly from the outset, and that many potential future difficulties are averted. MEDIN notes that the CCO applies a more prescriptive approach than BERR, for instance specifying the data formats that must be used.

The CCO approach has many attractions as it provides detailed instructions on how data should be prepared, and so allows little scope for ambiguity or confusion. However, the broad range of data

types and collection regimes across the marine sector is such that in many cases a single, detailed and proscriptive clause (such as that applied by the CCO) is not a practical option.

Therefore MEDIN proposes the use of a more generic data clause in tenders, given below, as a model. This does not preclude the application of a more detailed and specific clause, such as that used by the CCO, so long as the key issues are addressed.

Proposed Data Clause for use in Tender Specification:

MARINE DATA ACQUISITION

1 In all cases, standards applied to data collection and analysis as required in 2, 3 and 4 below shall be the highest that it is practical to attain and appropriate to the use to which they will be put.

2 Recognised standards must be applied by the Contractor (tenderer) and agreed by the Client to the process of data collection and processing.

3 Metadata must be provided with each data set in accordance with ISO 19115 or other recognised standard as may be approved by the Marine Environmental Data and Information Network. (see separate guidance for source).

4 The long term archival of data sets must be ensured by depositing the data in an appropriate Data Archive Centre (with any reasonable costs incurred to be met by the Contractor (Tenderer)) working to the standards established by the Marine Environmental Data and Information Network. (see separate guidance for source).

5 Ownership and copyright of data shall be agreed with the Client, and clearly stated in the contract.

6 The final report prepared by the contractor (tenderer) must include details about how this best practice has been undertaken and confirm that data have been submitted to the appropriate data archive centre.

Frequently asked questions about the Marine Data Acquisition clause.

Four questions are usually asked about this clause:

A) Why is this clause necessary?

Too many marine data have been lost in the past. There is a new initiative to ensure that appropriate marine data are submitted to data archive centres. This clause alerts potential contractors (tenderers), via the tender specification, to the best practice so that costs can be taken into account at the tender preparation stage. On acceptance of a tender, this then becomes a contractual commitment and a condition of payment. The tender documents could identify appropriate standards or sources of expertise to be referred to. MEDIN would be willing to supply advice and guidance.

B) How much work will be involved for the contractor (tenderer)?

The clause essentially enforces best practice, so additional effort should not be significant. Necessary effort from the contractor (tenderer) will usually involve: identifying suitable standards and /or engaging experts to assess standards; defining metadata and data standards and formats. The client will check that the metadata have been generated and are available; confirm that all data have been lodged in a Data Archive Centre. The contractor (tenderer) may be required to report on how they have adhered to these terms of the contract.

C) How will contractors (tenderers) know what to do?

By reference to appropriate standards and authorities. Guidance can be sought from MEDIN or appropriate agencies.

D) Does the clause apply to all marine data?

By default, yes. The only exception could be the requirement to make arrangements to archive the data with a DAC, - a special case would have to be made for any such exceptions.

FURTHER Guidance for individual clause conditions

Application and documentation of Standards

Evidence of application and documentation of agreed quality controls and other standards through the standard contract reporting mechanism will be a condition of payment.

Application and documentation of standards should represent normal “best practice”, and so should not result in any additional cost or effort for the contractor (tenderer).

Reference should be made to standards, protocols and recommended data formats documented by MEDIN, other appropriate sources (e.g. the UKMMAS protocols manual) or through reference to expertise as may reside within an appropriate agency or authority: SEPA, FRS, EA, CEFAS, JNCC, SNH, Natural England,…) if specific standards are not attached to this tender.

Generation and Publication of Metadata

Publication of verified metadata in an agreed format will be a condition of payment.

Generation of metadata is not onerous and should represent normal “best practice”. This therefore should not result in any additional cost or effort for the contractor (tenderer).

MEDIN will progressively publish guidelines and tools to support the creation of metadata in its recommended format. Contractors (tenderers) should consult with MEDIN DACs for advice on metadata content.

Metadata should be created and published for all data in all cases. The only potential, and rare, exception could be for reasons of security or possible commercial confidentiality. A specific case would have to be made.

Provision for long-term archival in a Data Archive Centre (DAC)

Proof that appropriate data have been lodged with a DAC will be a condition of payment.

This is likely to be the main source of additional cost and contractors (tenderers) should allow for this in their tender costs. The process for lodging the data would be agreed with the DAC in bi-lateral discussions.

All data should be lodged with a data archive centre unless the tender stipulates otherwise.

Ownership and copyright of data

Who would police /enforce this?

This would be a legal agreement between the contractor, the contracting body and the Data Archive Centre where the data are finally lodged. These terms would be enforceable throughout the life of the data.

Annex 3: Current Situation with Long-Term Core Funding for DASSH

DASSH has produced a short document summarising the capability and services it proposed to offer as its "Core" capability. The cost of this is £86,240 per annum.

The intention is to seek funding from a consortium of organisations, primarily government departments, who have a strategic requirement for a Data Archive Centre specialising on Marine Biodiversity Data.

Following the MEDIN Executive Meeting DASSH have been in direct discussions with Defra and Marine Scotland and are currently preparing a full document detailing costs, alongside key deliverables.