DRAFT

Northwest Environmental Data Network (NED)

Data Sharing Agreement[1]

Background

The goals and principles of making Pacific Northwest Environmental data openly and universally available have been defined in the Memorandum of Understanding on (see MOU, APPENDIX I).

The Participants who have signed the MOU have expressed their willingness to Memorandum of Understanding, which supports collaborative action and joint activities with respect to improving the collection, management, and sharing of environmental data and information.

Data providers often participate in several data sharing arrangements at different levels (thematic, community, regional, national, global).

NED data sharing should take place within a framework of due attribution.

Therefore, when registering their services with NED, the data providers agree as follows:

1. Data Sharing Agreements

1.1. Environmental data accessible via the NED network are openly and universally available to all users within the framework of the NED Data Use Agreement and with the terms and conditions that the data provider has identified in its metadata.

1.2. NED does not assert any intellectual property rights in the data that is made available through its network.

1.3. The data provider warrants that they have made the necessary agreements with the original owners of the data that it can make the data available through NED network.

1.4. The data provider makes reasonable efforts to ensure that the data they serve are accurate.

1.5. Responsibility regarding the restriction of access to sensitive data resides with the data provider.

---

[1] This data sharing agreement borrows significantly from the GBIF (Global Bio-Diversity Information Facility Information Sharing Agreement.

1.6. The data provider includes stable and unique identifier (SEE APPENDIX II) for an example about how the GBIF deals with unique identifiers) in their data so that the owner of the data is known and for other necessary purposes.

1.7. The NED Portal may cache a copy and serve full or partial data further to other users together with the terms and conditions for use set by the data provider. Queries of such data through the NED Portal are reported to the data provider.

1.8. Data providers are endorsed by a NED Portal Participant, if applicable, before their metadata is made available by the NED Portal Data Steward.

1.9. The NED Portal is not responsible for data content or the use of the data.

1.10. The NED Portal is not liable or responsible, nor are its employees or contractors, for the data contents; or for any loss, damage, claim, cost or expense however it may arise, from an inability to use the NED network.

2. Service Levels

NED Portal

2.1. Services provided by the NED Portal are managed in accordance with the NED Work Program.

2.2. The NED Portal's service provision includes software components and updates, interfaces, indexing and registry services, helpdesk, and training to assist the Participants to maintain Internet portals.

NED Participants

2.3. NED Portal Participants keep the NED Portal informed of their contact and service information.

2.4. NED Portal Participants maintain services that enable new and existing data providers in their domain to be integrated within NED Portal network, and the data owners be identified, as appropriate.

2.5 The NED Steering Committee is responsible for oversight of the NED Portal and creation of NED data Stewardship roles and responsibilities.

3. Definitions

- NED Participant: Signatory of the NED -establishing Memorandum of Understanding (MOU).

- NED Steering Committee: Administrative entity, empowered by signatories of the NED MOU to execute the NED Portal work plan, and maintain the central services for the NED Portal network.
- NED network: The infrastructure consisting of the central services of the NED Portal, Participant nodes and data providers. Making data available through NED network means registering and advertising the pertinent services via the NED central services.
- Node: A data provider designated by a NED Participant that maintains a stable computer gateway that makes data available through the NED network.
- Participant Node: An organizational unit designated by the NED Participant to coordinate activities in its domain. It may also provide data.
- Environmental Data: Primary and secondary data on Pacific Northwest species, water and habitats.
- Metadata: Data describing the attributes and combinations of environmental data.
- Data: Environmental data and metadata.
- Data sharing: The process of and agreements for making data freely and universally available on the Internet.
- Data provider: A custodian of data making it technically available. This may or may not be the data owner. If not they will have declared to NED that they have permission to make the data available.
- User: Anyone who uses the Internet to access data through the NED network.
- Owner of data: The legal entity possessing the right resulting from the act of creating a digital record. The record may be a product derived from another, possibly non-digital product, which may affect the right.
- Sensitive data: Any data that the data provider does not want to make available, e.g. precise localities of endangered species.

Version 1.0. NED Data Sharing Agreement (Based on GBIF agreement)

APPENDIX I

Memorandum of Understanding: Northwest Environmental Data-Network (NED)

This Memorandum of Understanding is entered into this the 12th day of October, 2004 between the:

NED Co-chairs:

The Northwest Regional Office of NOAA Fisheries (NOAAF-NWRO), as represented by the Regional Administrator NOAAF-NWRO, Mr. D. Robert Lohn

and,

The Northwest Fisheries Science Center of NOAA Fisheries (NWFSC-NOAAF) as represented by the Science Director, Dr. Usha Varanasi

and,

The Northwest Power Planning and Conservation Council (NPCC) as represented by the

Chairman, Judi Danielson

and, any other NED Co-chairs or Participants as described in Attachment A

(Hereinafter together referred to as the "Parties").

WHEREAS the NPCC has an overall planning responsibility for power and conservation issues arising from the development of the Federal Columbia River Power System (FCRPS) within the Columbia Basin.

WHEREAS NOAAF has a lead responsibility for anadromous fish conservation and management.

The Parties have agreed to undertake collaborative activities as detailed below:

I. PURPOSE

This Memorandum of Understanding (the "MOU") between the parties is intended to acknowledge and encourage discussion that may lead to further collaborative action and joint activities in areas of mutual interest with respect to improving the collection, management and sharing of environmental data and information.

It is understood that where our legal responsibilities require decision making based on good science and in many other areas our work depends on environmental data that is verifiable, of highly defined, high quality, accessible via internet technologies and is based on consistent or comparable methodologies and standards.

The MOU will serve as a basis for the development of mutually beneficial activities based on principles of cooperation, communication, trust, commitment and shared risk and benefit. It is a voluntary arrangement between the Parties and seeks to define the parameters of the relationship. It is understood by the Parties that subsidiary annexes, addenda or agreements will be required for all joint projects that may result from this MOU so as to ensure that the usual commercial concerns are duly resolved to the satisfaction of the Parties including but not limited to those involving financial and other commitments, as well as those involving intellectual property, copyright ownership and licensing of any existing or newly developed products.

For greater clarity, this MOU is not intended to, nor shall it, result in or create any binding legal obligations on the part of any Party and shall have no effect upon the governance or management of any Party.

It is essential that provinces, states, tribes, other federal agencies, local governments, citizens, and all interested parties have an opportunity to participate in the development of these information resources.

Under this MOU, the Parties seek to explore areas of mutual interest which may include:

- Identification of consistent data management standards and use of information technology tools to provide for an improved and more consistent data-network and exchange, including portals for public access.

- Capacity building, knowledge sharing and establishment of communities of practice in the areas of information technology and information management

- Coordination with existing data management entities and initiatives, for example with the Inter-Organizational Resource Information Coordinating Council (IRICC)

- Timely sharing of information system planning and plans for potential legislation

- Pursuing the use of consistent and common data collection and monitoring procedures, metadata standards, data dictionaries, data quality assurance and control, data cataloguing, and document management and promoting the support and implementation of the agreed upon standards

- Pursuing joint projects and co-funding opportunities with a focus to non-commercial, public uses

- Fostering common approach and communications for sharing data and information with other local, state, provincial, national, and international initiatives, especially through the use of open source applications and technologies

- Establishing linkages to science and technical information collections supporting sustainable resource management

- Any other areas or activities of mutual interest that may be deemed appropriate by the Parties including the development of similar MOUs to include other jurisdictions

II. ROLES AND RESPONSIBILITIES OF THE PARTIES

A. This MOU will be implemented by a Steering Committee to be established at the time of execution. An appointed designee from each signatory will serve as the lead representative for that Party on the Steering Committee. Parties signing as Co-chairs will lead the Steering Committee. Each signatory must appoint a designee within 30 days of signing the MOU.

B. The Steering Committee will be guided by this MOU and by corresponding Work Plans to be agreed on upon by the Steering Committee members.

C. Consistent with the enabling nature of this MOU the Parties will act to:

- Assemble working groups and other meetings as required
- Consider, and if needed, complete staff exchanges and joint training
- Exchange annual plans or master project lists
- Undertake collaborative projects and joint co-funding opportunities as appropriate, and
- Cooperate to make funding requests to further the purpose of this MOU

D. For specific projects undertaken related to this MOU, the Parties will commit key staff to work with one another on assigned projects as may be mutually agreed to in subsequent applicable memoranda of understanding or in addenda to this MOU.

III. PROCEDURES

A. At a minimum bi-annual meetings will be completed by the NED Steering Committee to review and assess progress with the MOU and measured benefits will be reported to the signatories of this MOU. All decisions will be made by consensus.

B. The Steering Committee, will as needed develop a proposed budget, required staffing level, and cost-sharing proposal for implementing this plan. The plan, budget and cost sharing arrangement shall be presented to each signatory for executive approval.

IV. ADDITION OF OTHER PARTIES

Upon agreement of the Steering Committee, other parties may formally participate in the MOU by signing ATTACHMENT A: Form for Adding Additional Participants or Co-Chairs to the Northwest Environmental Data-Network. All organizations involved in gathering and managing environmental data are encouraged to participate in the purpose and commitments described in this MOU.

V. TERM

The term of this MOU will be for 36 months effective 12, October, 2004 and ending 12 October, 2007 unless terminated sooner as per the conditions of this MOU. Furthermore, with the concurrence of The Parties, this MOU is renewable for subsequent terms. Prior to extension, a report shall be prepared by the steering committee describing the extent to which the activities performed under this MOA have accomplished their objectives.

VI. FINANCIAL COMMITMENTS

Signing of this MOU by the Parties does not, and shall not be deemed to impose on any Party any financial commitment to the other. Any financial commitment intended to bind any Party will first be approved by the Steering Committee, then reduced to writing and presented to the administrations of the Parties for their respective approval and subsequent mutual agreement.

VII. TERMINATION

Any Party may terminate its participation in this MOU at any time by giving the other Party(s) 60 days written notice to this effect.

SIGNED CO CHAIRS:

_____ _____10/12/04_____

Mr. D. Robert Lohn Date
Regional Director
NOAAF NWRO

Usha Varanasi

_____ _____10/12/04_____

Dr Usha Varanasi
Director
North West Fisheries Science Center

Date

_____ _____

Judi Danielson
Chair NPCC

Date

ATTACHMENT A: FORM for ADDING ADDITIONAL PARTICIPANTS or CO-CHAIRS to the Northwest Environmental Data Network (NED)

The undersigned agree to participate in the Northwest Environmental Data-Network.

Signer (Name and Organization) Date signed

Description of responsibility of the signer's organization

Signer (Name and Organization) Date signed

Description of responsibility of the signer's organization

Signer (Name and Organization) Date signed

Description of responsibility of the signer's organization

APPENDIX II How GBIF uniquely identifies data in the network.

How to identify data uniquely in GBIF network

GBIF Data Sharing Agreement and GBIF Data Use Agreement require that all data in shared through GBIF be identified uniquely. This is necessary for several purposes, like giving credit to the original providers of the data, linking digital data to the specimens in collections, and for removing duplicate data from analyses.

ABCD and Darwin Core, which are the data exchange standards used require as mandatory elements InstitutionCode, CollectionCode, and CatalogNumber. These form the basis for identifying data as described in the GBIF Data Architecure documents. However, data integration at global level, which is GBIF's mandate, requires additional considerations. This document describes how to use these and other identifiers so that they would scale up to a global level and for use across disciplines.

1. InstitutionCode

Definition from Darwin Core schema: *"A "standard" code identifier that identifies the institution to which the collection belongs. No global registry exists for assigning institutional codes. Use the code that is "standard" in your discipline."*

InstitutionCode is a short string, usually 3-6 character acronym, derived from the name of the Institution. There is no central registry that would guarantee uniqueness of InstitutionCode. Lists of such codes are maintained for particular communities by "coden providers". (A list of available coden providers is maintained at GBIF web site http://www.gbif.org/links/codon.) So, it is possible that the same code can exist in two different coden lists. Furthermore, the one institution can have several synonym InstitutionCodes in different communities.

In the DiGIR provider installation metadata (digir/admin/setup.php?show_provider_form=1), there is a field **Code**: This should contain the InstitutionCode, possibly with a namespace identifier indicating where the InstitutionCode has been registered (see below). It is not always clear why the InstitutionCode has to be in the provider metadata because it is supposed to be in the data. The reasons include at least the fact that often data records do not carry this mandatory element. Furthermore, the original source of the data can be recognised from the fact that the codes are the same in the data and metadata.

2. CollectionCode

Definition from Darwin Core schema: *"A unique alphanumeric value which identifies the collection within the institution."*

This is an abbreviation to identify individual collections within an institution. It is the responsibility of an institution to assign these and ensure their uniqueness.

The CollectionCode is also entered in two places in DiGIR provider metadata configuration. It is the first entry in Resource configuration, as **Code** which is defined as *"Unique code to identify the resource (usually the collection code)"*. It is also part of the **RecordIdentifier**, which is defined as *"An identifying code that can be recognized in the records of the resource. Often this is the combination of the hostCode and resourceCode. For example, ``MVZ Hild``"*.

The reason for including CollectionCode in both data and provider metadata are the same as with InstitutionCode.

3. CatalogNumber

Definition from Darwin Core schema: *"A unique alphanumeric value which identifies an individual record within the collection. It is recommended that this value provides a key by which the actual specimen can be identified. If the specimen has several items such as various types of preparation, this value should identify the individual component of the specimen."*

4. CodenProviderCode

This is a new code. Because the same InstitutionCode can exist in multiple coden lists, and data will be integrated across multiple disciplinies by global portals, like GBIF, it will be necessary to identify from which community the data originates. There are some alternatives how to do this. One alternative is that the coden providers get together and harmonise their data. Another alternative might be that GBIF Participants harmonise the codes within their domain, i.e., within countries. However, it is mainly in the interest of GBIF as a global integrator to guarantee uniqueness of data identifiers at global level and across disciplinies, and probably a solution can be built on this basis.

Therefore, GBIF would maintain a list of coden providers it recognises and maintain the CodenProviderCodes for these. Where needed, it can be used in front of the InstitutionCode to identify the origin of the latter.

Furthermore, to build a fully qualified combination of the codes, a designation of the network, i.e., GBIF would be prepended. This is necessary to identify who has issued the CodenProviderCode. Therefore a fully qualified identifier of data has the format
`NetworkName: CodenProviderCode: InstitutionCode: CollectionCode: CatalogNumber`
Example
`GBIF:IH:S:BRYOPHYTA:12345`
which would stand for
GBIF Network, Index Herbariorum, Swedish Museum of Natural History, Moss collection, record number 12345.

5. Recommendations for using the codes

5.1. Use a locally unique identifier

It would be too much to ask every data provider to use for all data a globally unique identifier, i.e., the fully qualified combination in canonical form as described above. Instead, a "subsidiarity principle" would be used. Records in databases are required only to use as many Codes as it needs for identifying data uniquely locally, which means at minimum the CatalogNumber. If a database (resource) has data from several collections and institutions, these would have to be found in data but not otherwise. In other words, these identifiers can be in the provider and resource metadata, if they are the same for all records.

Under what CodenProviderCode a data provider installation falls, would have to be found out when registering the provider.

5.2. Lookup services

In order to facilitate inclusion of new data providers in GBIF network and help them to choose the correct identifiers from start, some lookup services could be envisaged.

Not all institutions are aware that they should register with the coden providers. Therefore it is recommended that the GBIF Participant Nodes maintain a list of institutions in their domain. They could proactively work with the coden providers to establish codes for the potential data providers in their domain. It would be good if such lists of institutions and their codes could be available on Participant node web sites. This list should be a simple table formatted with the following columns

1. InstitutionCode: 3-6 character acronym
2. InstitutionName: Full name of the institution in English and other languages
3. URL: Main web page of the institution
4. PostalAddress:
5. CodenProviderCode: Any Coden Providers where this code is listed. This is either the abbreviation of the GBIF Participant Node itself or abbreviation of a coden provider, or both, separated by commas. See the annex. The first listed Coden Provider is recommended to be used.

Several lines should be inserted in the table in case one institution has several codes. The web page should also give instructions how a new institution can get listed at the appropriate coden provider.

At some point GBIF Portal could syndicate all the tables from nodes and coden providers and create a global lookup table. This would be used by provider software installation packages and UDDI registration scripts.

5.3. Portals remove duplicates

Sometimes data custodians hesitate to make data available, because they know that the data is also present elsewhere. Indeed, it is very likely that same data records will become available from several data providers, and this should not be seen as problem, if the data is identified appropriately as described above. Many kinds of analyses do not even require that duplicate data be removed.

Sometimes data is updated or additional fields are added. If there are multiple copies of the data, they become out of sync. The field DateLastModified should be used consistently to help recognise such cases.

When needed, the original provider should be possible to recognise from the fact that its DiGIR provider metadata is identical to the values in the data. GBIF Portal will primarily serve data from these original sources. Other portals may choose other strategies.

Probably data duplication cannot be prevented entirely, but with a consistent management of data identifiers the problem can be addressed, at least with some effort.

5.4. Change of data from institution to another

Sometimes collections change names and may even move to other institutions. If this happens, it is suggested that the original data is kept online for a period of 4 months, and all other fields made blank except the above three fields. Naturally, DateLastModified should be set to current date. In addition, RelatedCatalogItem should get the new RecordIdentifier. This will effectively remove the data from caches of the portals. See the definitions in Darwin Core schema for further explanation.

5.5. Recognition service of data providers and data ownership

A lookup service for data ownership could be envisaged. At some point GBIF Portal may provide a function which would allow a user to upload a Darwin Core formatted file and generate a list of institution names and collection names which should be recognised in publication acknowledgments.

5.6. Formatting of data records in databases

Each Darwin Core and ABCD record carries the three codes, but not CodenProviderCode nor NetworkName. It is recommended not to include these in the data records, mainly to keep them short. A provider may serve data also to other networks than GBIF and identifiers specific to one network only would not be desirable.

5.7. Use of UDDI and UUID

GBIF UDDI registry has universally unique identifiers (UUID) for all objects it deals with, including "Businesses" and "Services". These roughly map to institutions and their data provider software installations, respectively. These identifiers are used internally by portals to keep track

of the data providers, but as they are not easily readable by humans, they are not recommended for wider use at this point.

5.8. Coden Providers and GBIF UDDI registry

GBIF will probably have to recognise some Coden Providers to get the InstitutionCodes organised. Choosing them is not necessarily an easy task. Only very few Coden Providers enjoy global recognition. Basically, a Coden Provider will have to be reachable fast and offer a stable public service that allows finding out what a code means. In advanced form, XML lookup tables could be offered. There are also directories of institutions that do not maintain any codes for their entries.

There is a possible link with UDDI and Coden Providers. Provision of codes can possibly be a kind of an "authentication" service that could be registered in UDDI. Let's explore this for a while.

A data provider can serve data originating from several communities. That means that a provider can have several InstitutionCodes in its data, and several could also be relevant for metadata. So, registering a CodenProviderCode as identifier of an institution in the GBIF UDDI Registry makes little sense.

A resource roughly corresponds to a collection. This would allow associating resources with Coden Providers of particular communities. Basically, it is the "ResourceIdentifier" in DiGIR provider metadata, that would have to carry also the CodenProviderCode. There is no place for this in GBIF UDDI Registry, because it is only "businesses" (institutions) and their "services" (data provider applications), and not resources that are registered there.

In summary, unfortunately there is no direct way of identifying a namespace, like a CodenProvider, with a UDDI registry entry.

6. Discussion

This is a live document. Minor technical comments should be sent to the address below. Other comments should be sent to the mailing list of GBIF Nodes <nodes.gbif@ig.circa.gbif.net>, where also earlier discussion on this subject has taken place.

Annex

Codes of Coden Providers are given in square brackets.

See http://www.gbif.org/links/codon/

Ichthyology [ICH]
Index Herbariorum [IH]

Insect and Spider Collections of the World [ISC]
International Directory of Botanical Gardens [BOT]
Malacology [MAL]
Mammal Collections in the Western Hemisphere [MAM]
Microbial collections [MIC]

GBIF Participant Nodes are also coden providers in their domain. See http://www.gbif.org/links/codon/particodes Syntax example: GBIF-FI for Finland, GBIF-ITIS for ITIS.

Also see
Zentralregister biologischer Forschungssammlungen [ZEFOD]

Version 0.4.1 Draft Hannu Saarenmaa 2004-1-22 18:02