2003 ABCD comments commented – January 2004

While working on the update of ABCD, I have collected my notes on the comments provided by other people and included them in this document. Please note that some of the comments refer to earlier versions of ABCD, so that they may have been implemented already e.g. in versions 1.10 or 1.20 of the schema.

Comments from: Andrea Hahn (AHA), Anton Güntsch (AGU) , Charles Copp (CCO), Donald Hobern (DHO), Dave Watts (DWA), Hannu Saarenmaa (HSA), Markus Döring (MDO), Neil Thomson (NTH), and Peter Strobl (PST)

Jan. 29, 2004. Walter Berendsohn (WGB)

1. Dataset – general and metadata issues

1.01. NTH: A number of the new elements that have been put in have taken the default of "mandatory". An example is <ContinentOrOcean> so a review of optionality would be good.
WGB: Done.

1.02. HSA: Shouldn't "OriginalSource" be under Datasets, not Dataset? One can derive many Datasets from one source. It is unlikely that one XML file would contain datasets from many sources.
WGB: I do not agree. For example, the SysTax database in Germany will soon be accessible and it will supply a single ABCD document as a response when asked, e.g., for Apis. The data, however, will come from several entomological datasets hosted by SysTax, and those are from several institutions, each of which is an OriginalSource.

1.03. HSA: GatheringCode is good. However, ObservationUnitIdentifier which serves similar purpose is named differently. We need to be consistent how we call them (ID, Code, GUID, ...) We also need similar IDs for other elements such as Site, Person, ... This is necessary so that we can go back to original data if needed, or locate it elsewhere.
WGB: I think this is an important point for further discussion and standard development. I agree that clearly defined key elements should be named accordingly. Currently, this refers to SourceInstitutionID (changed from SourceInstitutionCode), SourceID (changed from SourceName) and UnitID (unchanged) in ABCD, which are roughly equivalent to the three IDs in the Darwin Core. In renaming, I followed HSA's suggestion for these elements. However, it has to be pointed out that these have provider-determined contents - the only rules are that (i) the SourceInstitutionID should be unique globally (it may be a code or a name or a combination of code-source and code [e.g. IH-K]; (ii) the SourceID must be unique within the institution, and (iii) the UnitID must be unique within the source. The naming of other ID's, Codes and Identifiers used in the schema now refers to content rather than to their role in the processing of the information. I posit that we should keep it at that as long as no generally agreed scheme for ID's in the biodiversity information emerges. ObservationUnitIdentifier is a good example: it may or may not be the unit level ID - this is left to the data provider and their domain-specific needs. However, the UnitID must be given, if decided so, it can be the same as the OberservationUnitIdentifier.

1.04. NTH: <DatasetDerivation> - is there a need for this to be repeatable? Our feeling was that it shouldn't be if they have come from one supplier.
WGB: Yes, it needs to be. For example: NHM in Karlsruhe (original source) provides Butterfly data to SysTax in the German GBIF project; SysTax provides data to the Global Butterfly Information System (first derivation), which may be accessed by a GBIF portal, which in turn provides a secondary derivation. I think we discussed this extensively in Singapore.

1.05. NTH: I do not remember why repeatable elements were now to be wrapped in a container element. Was this some kind of good practice technique - or does it just bloat the element count? Either way, there is inconsistency, with some elements wrapped and others not (yet) wrapped.
WGB: Done (I hope). The wrapper elements make all the repeated elements in a document based on the schema addressable under a single path.

1.06. NTH: .. would like a flattened version with an index number against each element ....
WGB: We discussed the question of indexing of elements (and attributes), especially with regard to the simplification of version changes (automatic mapping and re-writing of configuration files and interfaces, etc.). However, we concluded that in any case a mapping old-to-new has to be provided, and that unique identification of elements in types would be not be helpful for version changes anyway. Thus, we should use the individual xpath as the index.

1.07. NTH: As BioCASE is no longer tied to DiGIR, it follows that it is no longer tied to Darwin Core - although this would remain a useful option. Should we determine a smaller subset of access points? Maybe based on the Who / What / Where / When questions?
WGB: Yes, this would indeed be very useful. It should be part of the forthcoming documentation of ABCD, for which all of us lack the time. In Oeiras we already defined on of the name elements (the fully concatenated one) as an access point, same as the InformalName.

1.08. MDO: All elements should have a type.
WGB: Done; MDO wrote a program to test for untyped “leaf” elements.

1.09. MDO: Think about using an optional ID attribute for every repeatable element to allow for better imports (you can detect real duplicates with this - so you know Müller ID=1 is really not the same record as Müller ID=2 within the same dataset). This would be particularly useful for data exchange purposes (import of dataset).
WGB: I would like to put this up for further discussion. **

1.10. SBL: .. narrow the definitions of all these string elements to something more precise than "xs:string". Most are probably xs:Name or xs:NameToken or something like that. The O'Reilly XML Schema book has a pretty good discussion of simple XML data types.
WGB: We think that we should see how things develop first. One of the principles of ABCD is not to be too prescriptive. This puts more of a burden on the developers, of course, but a bunch of invalid documents because of to tight typing would not help us either.

1.11. PST: Since the file is rather large, would not it be better to split it up into senseful pieces, as XML schema offers several options for modularization? I also guess this would improve readability and ease of maintainance. Or vice versa, are there any technical problems hindering this?
WGB: The version for download on the web represented a file concatenate from several subunits (as depicted on the SourceForge site). For this the Includer software provided by Markus Döring (also on the SourceForge site) was used. While the old subdivision into modules largely followed editorial puposes, a more meaningful distribution is provided in versions >1.20. Also, a zip-file with all individual files will be provided besides the concatenated version. The latter is needed for the Schema Viewer to work.

1.12. WGB: Remove OpenIDType and replace occurrances with xs:string.
CCO: No objections, was introduced earlier when the thinking was still more in modelling terms.

1.13. DWA: I note there is not a lot of controlled vocabularies with some of the ABCD elements. I presume that will increase with time.

WGB: See discussion about that subject during the Oeiras meeting (http://www.bgbm.org/TDWG/CODATA/OeirasWorkshop.htm). In some cases controlled vocabularies were introduced (especially in the domain specific parts of the schema, where standards are used). In other cases it will be the work of the providers acting together with specific user communities to ensure that data adheres to certain (sometimes community-specific) term lists.

1.14. PST: Formatting the code using indents would be rather nice for people like me who look at the code directly.
WGB: This has been corrected, the code needs to be imported into a tool like XML-Spy once after it has been concatenated with the Includer-Tool.

1.15. PST: It seems the file was not created using a standard text editor since there are a lot of tags only with whitespace content. In this place may I warn you not to believe any of these tools creates valid XML schema, but rather use an external validation tool.

WGB: Whitespace content was removed and the schema should be valid.

2. General unit-level data (incl. associations, images, etc.)

2.01. NTH: <UnitID Numeric> - should be alphanumeric rather than decimal.
WGB: We introduced this element in addition to UnitID to specifically to allow searches on a range of numbers (where possible). Essentially, it should be a duplication of (the numeric part of) the UnitID. This wouldn't make sense if it's alphanumeric.

2.02. NTH: <UnitAssociation> - text says that the association should be with another unit in this dataset, but the implication of the structure and the preference of the TC meeting (plus logic) is that it should be possible to go across datasets.
WGB: Done. I also deleted the <Rule> saying the same.

2.03. MDO: UnitDigitalImages/UnitDigitalImage/ImageSize should mean the pixel size of a picture. I would reuse ImageSize as this: (1600x1200 pixel) and add an element filesize for the amount of data (12,4 MB)

WGB: Agreed and changed.

2.04. MDO: A remark for UnitID should be added not to use primary keys of a table for this, as it should be assured that the unit IDs keep stable when importing/updating databases.
WGB: Done

3. Identification/Taxon issues

3.01. DHO: We are still using TaxonIdentified as well as Identification. If the Synecology element is staying it should use Identification.
WGB: I don't think so. The synecology element records observations taken at the time of the gathering event of a specific unit, so the identification event data should here be the same as the gathering event data. Using formal identifications at this place would unnecessarily increase nesting. For direct synecological observations, individual units would be created for each of the observed taxa, so in this case the identification type would be used.

3.02. DHO: I think that FullNameAuthorYear [changed to FullScientificNameString] is now intended to be the basic location for entering a scientific name. It is after all mandatory. Are users supposed to be able to use this field to enter whatever they have as the best available identification ("A-us b-us Author 1972", "A-us b-us Author", "A-us b-us", "A-us sp.", "A-idae sp.", etc.), even if it is not complete?
WGB: Yes; I hope the new naming makes this easier to recognize.

3.03. AGU: The Complex type given under /DataSets/DataSet/Units/Identifications/Identification/ScientificName/ScientificNameAtomized/Zoological/CombinationAuthorTeam does not have children.
WGB: I think this was solved for v. 1.20 already.

3.04. HSA: Identification should have a person (ID) and timestamp.
WGB: Identification has an Identifier (which may be a person, another [legal] body, and/or a reference from which the entire event was taken. With respect to PersonID see comment under 1.03.

3.05. HSA: Maybe [Identification should have] also some attribute giving the certainty % by which the identification was done to that level (example 100% to genus, 50% to species).
WGB: I would like to refer this to the groups handling due process and best practice in collection digitization. I personally doubt that we can introduce anything numerical here. In Botany, there is a tradition of expressing a certain doubt in an identification by using, e.g., the abbreviation “cf.” (which is accommodated in the IdentificationQualifier element).

3.06. HSA: Identification should also have the method by which it was done.
WGB: Up to now, only the IdentificationNotes could be used to indicate, e.g., that molecular or acoustic methods were used for the identification. Especially in observation records, this can be an essential element for data quality assessment. Again, there is room for standardisation, for the time being, I added a simple text field.

3.07. HSA: I don't think the simple Identification History element ... is useful.
WGB: I came across databases, which dump the preceding determinations in a text field once a new identification was made. As in many cases, we'd prefer the structured version, but we want to make sure to get the data even if it's not properly structured.

3.08. HSA: In Identification InformalName Language there should be a provision for giving something like Taxonomic Serial Number of ITIS. I don't think TSN is just a "language" as the current schema would accept it. If such number or code is given, the namespace should be given as well.
WGB: An identification should always reflect the result of an identification event (i.e. a decision taken at a certain time and by somebody). Linking the result to an external name-based system like ITIS runs the risk of changing the actual results (e.g. by later changes in the taxon circumscription given by those systems). I think this is a good suggestion for the future, when concept based taxonomic services become available on the network.

3.09. HSA: Indeed this need for being more specific on namespace that "NomenclaturalCode=" applies for ScientificName as well. This should point to some Taxonomic Name Service with closed set of names and URIs.
WGB: There are two issues at hand here. (i) The respective subtype of NameAtomized indicates the fact that an atomized name is structurally handled according to a specific code of nomenclature. Keep in mind that issues like the validity of a name, publication, synonymy, homonymy etc. don't refer to the results of identifications of collection units but are to be treated in the context of taxonomic and nomenclatural systems. (ii) The taxonomic domain for any identification (also those containing only a scientific name string or even an informal name) can be indicated by using the HigherTaxon element. This is recommended to simplify searches.

3.10. NTH: <InformalNameString> - should be repeatable.
WGB: Look at the new name subtype. Different names should ideally be the result of different events.

3.11. MDO: A controlled vocabulary should be provided for higher taxon ranks.
WGB: Done.

3.12. MDO (following the subgroup discussion in Oeiras): Incorporate DHO's new IdentificationType structure to make the concatenated name mandatory.
WGB: Done.

3.13. DHO: During the WFCC meeting, we discussed mappings between their standard data sets and the ABCD. Most things I understood, but I could find no ABCD element in which to encode the host taxon or substrate from which the microorganism was collected. Is this supposed to be part of the GatheringEvent? How would you expect such taxonomic identifications to be included? Surely not just buried in Synecology?
WGB: Your question clearly shows that we should have had a microbiologist in Singapore, I think we eliminated the two elements needed to cover substrates during that meeting.
In fact, this is a wider question relating to directed relationships between objects (virtual or real) in the dataset. This includes also questions like host/parasite relationships, organisms found in the stomach content of predators etc.
There are several cases here: (1) We have two units which are related to each other. (2) We have a single unit consisting of two organisms that are related to each other. (3) We have a single unit consisting of one organism and its substrate, which is (part of) the gathering site [e.g. a type of surface of non organismic character, rotten wood of unknown origin, a cellar wall, etc.).
Case 1 is covered by the UnitAssociationType where the first three elements identify the second unit (e.g. the sample of the substrate organism) and the AssociationType defines the relationship (e.g. "grows on" or “isolated from”).
Case 2 may, theoretically, be resolved into case 1 by actually creating two derived units from the single one (taxonomic homogeneity principle in the BioCISE model - the moment you define a new, previously undefined taxon in your mixed sample, you effectively create two new derived units in your database).
However, in practice we will continue to have two qualified identifications of the same unit. I looked all over ABCD, this is something that was there and is now missing. I think it was the attribute "SpecialTargetCategory" of "Identification" which we did away with in Singapore because we couldn't find out what it meant. So this has to go in again, perhaps Role would be a better name? At a later stage we probably need to provide a controlled (but extendible) vocabulary.
As a result, the host / parasite relationship (where there is only 1 unit) would turn into two preferred identifications, one of which has Role="host", the other Role="parasite". Another example would include “substrate” and “isolated strain”.
This is the clean solution, because it allows to involve the identifier of the host (eg. a plant) as well as the identifier of the organism using it as a substrate. I realize that many databases do not have that information, but they should be encouraged to provide it.
Case 3 was originally covered by the GatheringSite, but it isn't any more because we actually restricted it to the geographical/ecological side of things. I first thought to propose to include a new element to UnitCollectionDomain/CultureCollectionUnit/ (e.g. SubstrateMaterial).
However, I keep coming back to CDEFD and BioCISE modelling results and I believe Gregor Hagedorn first brought up the thought of treating such matters as non-taxonomic identifications. Sometimes considerable effort is made to provide a good description or identification of the substrate. This would also open up the way for the inclusion of identifications of other collection objects (e.g. minerals) that are housed side by side with biological objects in natural history museums. Consequently, I have simply extended the IdentificationType with an element MaterialIdentified as an alternative for TaxonIdentified.

4. Contact / Person / Organisation issues

4.01. DHO: I think we all agreed to remove SortPersonName from PersonName, and therefore to replace PersonNameType with a plain string.
WGB: Well, this was the editorial meeting's opinion. However, I got other suggestions later that support a separate type for person names (e.g. HSA wants a possibility to include an ID here, and that does make sense if we make progress on a 'Naturalists Directory' as envisioned in the SYNTHESYS project). I agree that we should rename the SortPersonName element, because it is really much more important for queries than for sorting purposes. The other element is necessary to maintain original text where wanted. In version 1.3 I called them PersonName and PersonNameLastFirst.

4.02. HSA: "PersonName" contains a similarly named element. The containing element better be called "Person".
WGB: The type has been changed accordingly (already in v. 1.2)

4.03. HSA: Person should have an element "PersonCode" that would support any person identification schemes and remote namespaces. Using such identifier, all the person data that is repeated could optionally be retrieved from a directory server somewhere. Keeping track of and certying observers is important if there is no specimen to go back to for verification.
WGB: I put this suggestion in the annotation of Person, it should be implemented once such directory servers become operational [we are waiting for a GBIF initiative here].

4.04. HSA: Person should have a public key. In case someone verifies an identification, there must be a possibility to leave a digital signature with the identification.
WGB: Another suggestion that I think is a little ahead of current usage, but I have included it in the editorial comments of the PersonType, too.

4.05. HSA: Related to above "OrganisationCode" should have a qualifier to identify a namespace, for instance a directory server.
WGB: Another suggestion that I think is a little ahead of current usage, but I have included in the editorial comments of the PersonType, too.

5. Gathering event

5.01. AGU: Collectors are represented in ABCD with an unbounded Element GatheringAgent. Alternative text representation is missing.
WGB: Has now been included. Should we make this obligatory and use it as the search access point? This would than work in analogy to the taxon name, only that a substring search should be used because we don't have a Code for person teams.

5.02. HSA: "DateTime" under Gathering should have a qualifier available to designate whether the element concerns Period="Begin" or Period="End". It is true that this could be inferred from date but it cannot be inferred from time.
WGB: I think this is (now?) fully covered by the current DateTimeType in ABCD. The *Begin elements and *End elements for ISODateTime as well as TimeOfDay (in combination with either a ISODateTime expressing only date or a JulianDay) fully specify the period.

5.03. NTH: There is some duplication in <GatheringEvent> now that the <DateTimeType> includes elements such as <Calendar>
WGB: I suppose that this had been solved already in version 1.20?

6. Gathering site

6.01. AHA: The element <SiteText> (free-text description of collection site) does not exist any more in ABCD 1.01. Since site information very often just comes as a text field in provider databases, I think we need this element. As a fallback-option the element <LocalityText> could be abused, but this is dangerous (since it is intended to just carry the original label information).
WGB: You are right about not using LocalityText to further specify the collection site, this is for the entire label text as far as it concerns the collection site. I think the element AreaDetail should be used for free text descriptions of the site where these are given in addition to the atomized (higher-level) elements such as country, nearest named place, etc.

6.02. NTH: Many of the separate elements under <GatheringSite> which is by far the most unwieldy part of the schema, could be eliminated in favour of using either <LookupType> or <MeasurementType> which is what many of them actually are anyway. By using these generic structures, the element count could be reduced and the flexibility enhanced. For example, <Altitude> is a just a measurement.
WGB: It seems that this was done already in version 1.2, or is something missing?

6.03. NTH: For observations, there may need to be a bit more descriptive material. Steve Wilkinson will propose a structure which I will forward on receipt. He felt that what was currently in <Biotope> is not sufficient.
WGB: I am looking forward to this, and we are in direct communication with the group in the course of connecting their observation warehouse system to BioCASE.

6.04. AGU: Change ISO2Letter to ISOCountry and allow entry of the 4-letter ISO3166-3 codes for formerly used country codes.
WGB: Changed Element name to ISO3166Code and deleted ISO3Letter (is also defined in ISO3166-1).

6.05. DWA: you have the examples of the 2 letter and 3-letter country codes mixed up.
WGB: Should be all right now.

6.06. SBL [during Oeiras meeting]: Gathering biotope measurement is time dependent so it should be under event.

WGB: The elements under GatheringSite are all more or less time-dependent. Countries change their borders, named places change their names, etc. It is very difficult if not impossible to actually define the borderline. I'd thus prefer to keep things as they are until people using these elements provide more input (and until we perhaps replace many parts of GatheringSite with parts of other schemas).

7. DateTime

7.01. MDO: this is the new DateTimeType for ABCD. The regular expression checks days etc. but allows 31 days for all months, because it would otherwise become unwieldy.
WGB: Incorporated.

7.02: [Oeiras meeting]: We need to add an Explicit field for ranges of date or time to indicate that the event actually took place for the time of the range given, instead of at some point in time during that period.
WGB: Added element PeriodExplicit to DateTimeType.

8. Measurement and Fact types

8.01. AHA: under MeasurementAtomized/ParameterMeasured, an information is requested that, imo, is already unambiguously defined by the position of the type within the document (e.g., altitude). Therefore, it seems a bit redundant, unless the type shall also be regarded out of context.
WGB: I agree that it normally can be ignored, but this type could be generally useful.

8.02. HSA: Measurement should support discrete values. How does one represent, for instance, that an observation concerns Gender=Male, Generation=3, or Stage=Larva?
WGB: UnitMeasurement covers any character/character state combination with a numerical character state. To take your example:
ParameterMeasured = Generation; MeasurementLowerValue = 3.
UnitFact covers textual charater states. E.g.:
FactType=Gender; FactText=Male. FactType=Stage; FactText=Larva.
I have tried to make that clearer in the annotation given for the respective elements.
[N.B.: Recognizing their importance, both gender and stage are covered by the ZoologicalUnit subtype of the UnitCollectionDomain section in the schema (Elements ZoologySex and ZoologyPhase).]

8.03. NTH: The values in <MeasurementType> should have the ability to accept text. It is unlikely that this element will be used for sorting purposes, but some values may be descriptive.
WGB: See answer to HSA above. In short: UnitMeasurement is for numeric results, UnitFact for textual results (character states or free text). Tried to make that clearer in the annotation text now.

8.04. HSA: For brevity, measurement should support in addition to lower and upper value, just "value".
WGB: I disagree. Searches are made easier with a single element for either lower and the only value.

8.05. MDO: A controlled vocabulary should be provided for Measurement units
WGB: This should be discussed. **