ABCD comments commented – June 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. Some are rather trivial, but I did not edit them out to show that I pay attention ... 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.

Unresolved issues are marked by ** and highlighted in red.

Comments from: Adrian Rissone (ARI), Agnes Kirchhoff (AKI), Alexander Kroupa (AKR), Andrea Hahn (AHA), Anton Güntsch (AGU), Ben Richardson (BRI), Bob Morris (BMO), Charles Copp (CCO), Dave Watts (DWA), Donald Hobern (DHO), Gregor Hagedorn (GHA), Hannu Saarenmaa (HSA), Jessie Kennedy (JKE), John Wieczorek (JWI), Markus Döring (MDO), Mark Jackson (MJA), Michael Malicky (MMA), Neil Thomson (NTH), Patricia Mergen reporting for the Belgium Coordinated Collection of Microorganisms (PME), Peter Strobl (PST), Sabine Roscher (SRO), BioCASE Technical Committee (TC), and Yde de Jong (YDJ)

June. 30, 2004. Walter G. Berendsohn (WGB)

Contents

1. Dataset – general and metadata issues

** Dataset-level metadata will be completely revised to provide a common structure with the other TDWG standards (SDD and Names/Concepts).

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: A dbid attribute added to every repeatable element would be possible and is currently under 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.

1.16. SRO: Thinking about the PlantGeneticResources elements I miss an indication whether a DataSet will also be provided by another network, e.g. the EURISO-network. This might help to avoid redundancies in the sense that the portal can filter or reject data from the EURISCO-network that are directly delivered by the 'local' data provider or the other way round. (I remember that the BioCASE metadata schema has an element called network.)
WGB: I added the element OtherNetworks/OtherNetwork to cover this – it's the ID in the UDDI Registry (the UUID) for that provider.

1.17. GHA: The Identifier of a dataset refers to the institution and collection holding a specimen (or a unit dataset). However, databases exist (e.g. within the GLOPP project) that unite first digitizations of specimens from several collections. Moreover, these data may be highly processed and quite different from the data originally taken from the labels in the collections.
WGB: SourceInstitutionID and SourceID were moved to the UnitType. They have thus to be repeated for every individual unit, but this makes it possible to have a dataset containing units from more than one original source. Documentation of further provision, derivation, or audit trail details of data remain on the level of the dataset.

1.18. GHA, BMO, WGB: The rights and statements types should be united and cleaned up. All statements should be derived from a generalized statement type.
WGB: Done.

1.19. NTH: Before you roll out v1.4, you may wish to check out an apparent anomaly with the ContactType. It appears that there are currently two versions being used in the Schema. There is the condensed version used e.g. in Dataset / DatsetDerivations / Supplier and there is the older, extended version that still pops up in e.g.

UnitDigitalImages / ImageIPR / LegalOwner.
WGB: Solved.

1.20. MJA: [Using ABCD ..] the resulting file has to carry a lot of higher tags which are meaningless in the [my project's] context, and many of the tags have names which fail to convey the necessary meaning. A good example is <GatheringAgents><GatheringAgent><AgentText>Sacleux</AgentText></GatheringAgent></GatheringAgents> which I can more simply record as <Collectors>Sacleux</Collectors>. I've concocted a simpler schema .. [and] tried to use element names from the ABCD namespace, but found that many of them sounded awkward in this new context so I dropped that idea. I've tried to keep the semantics and format of the fields the same. I'll have a go at building an XSL to transform the output from this to ABCD, though I guess this is more of a theoretical exercise than anything else.
WGB: This is of course the problem of every individual project, sub-discipline, special interest group, etc. using ABCD. The idea is to provide a generalised schema for all biological collections, which can be used (and be cut down) for many purposes. I should think that actually providing ABCD from your database with an XSL transformation to the simpler format to provide for the project is the better way to go. However, it is already a very important step forward if you keep semantics and format of the elements the same. If you go that way, please do identify this in the annotation of the elements in your schema (and please cite the version of ABCD you are referring to - we are working on a concept-tracking mechanism for different versions and standards in collection data).

1.21. MJA: the SourceLastUpdated field is mandatory but redundant

WGB: The last update can indeed be recorded for the entire dataset or for individual units. However, the unit-level one cannot be mandatory, because most databases can't specify it (although, where present it can be extremely useful). These can at least try and set the one on the dataset level. This is useful in any case, because if no change is indicated, indexing mechanisms do not have to search the individual unit records.

1.22. MDO: The distinction between OriginalSource & DatasetDerivation is unclear to nearly everyone I talked to. What about a clearer seperation between the supplier and the derivations (which may still include a supplier): OriginalSource (not repeatable); Supplier (not repeatable); Derivations (optional, repeatable)?
WGB: The entire structure to describe Datasets was changed and is still being discussed with the SDD (Structure of descriptive data) standard group. SDD has taken on some of ABCD structures (e.g. the Datasets – Dataset construct), ABCD will take over SDD's TransformationHistory construct. There are several questions still under discussion.

1.23. OriginalSource/SourceWebAddress should be named similar to the rest: OriginalSource/SourceURL.
WGB: All URLs where changed to URIs, and WebAddress does not exist any more.

1.24. GHA, JKE: Common rules for attributes and element names: element names with capitalised word elements, attributes all lowercase.
WGB: accepted, change in progress.

1.25. DHO: I have noticed a couple of minor glitches in the ABCD Schema naming .. There is no consistency overall between using "-ise"/"-isation" (UK spelling, e.g.

“Organisation”, “Mineralisation”) and “-ize”/”-ization” (US spelling, e.g. “Organization”, “Atomized”) for element names. The most obvious case is: GatheringAgent/Organization/OrganisationName.
WGB: solved, used British spelling throughout (exception: xs: types).

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

2.05. MJA: there are no fields for some image metadata which we will probably have to add in (Image Resolution, Capture Equipment, Date Image Created and Image Creator)
WGB: These have been added to the ImageType, as well as some other items that were identified in the recent ENBI/GBIF workshop.

2.06. PME: Accession Number: Use **SourceInstitutionID**, **SourceID** and **UnitID**, for UnitID not just the numerical part but also the Acronym which is a complete part of the record unique identifier, ie LMG 115 and not just 115. This is for example important as for DMZD the letters of the strain ID is not DMZD but only DMZ. Works are ongoing on a Global Unique Identifier for each strain, but not yet in use.
WGB: Use DSMZ – DSMZ – DMZ115 for the three mandatory ID’s.

2.07. PME: Other culture collection numbers. These are not only previous IDs but contemporary strains that are kept in other collections under a different Identifiers. Suggestions would be to put them in **AssociatedUnitID** with mention of strain duplicates in the **AssociationType**.
WGB: I don’t think the AssociationType is the place to put this. If the history (past and present) is used properly, all the data can be stored there (there is a date for PreviousUnit). If the collection wants to record strains derived from this unit, they should store that new unit in their database or access it by way of their unit id found in the recipients database in the SpecimenUnitHistory.

2.08. PME: History of deposit: The order is important, is there a possibility to use a nested structure?
WGB: The history can either be stored as a text under UnitStateDomain\SpecimenUnit\SpecimenUnitHistory\PreviousUnitsText or as a sequence of references to other units (given by the three-partite ID) under ...\previous units. If this unit is not available from the original holding institution, it can also be stored in the present institution's database.

2.09. PME: History of deposit: Important is also to mention the Name under which the strain is known during the history and to add the names and contacts of the Depositor and Isolator. Is it possible to add a **ContactType** in **SpecimenUnitHistory** where the role of the contacts can be mentioned as for the supplier in the DatasetDerivation part WGB: If all this is not stored as a single text (see 2.08), the depositor and date is stored with the individual unit under SpecimenUnit/Acquisition/**AcquiredFrom** (a ContactType), where a date can be stored, too. The isolator is stored under SpecimenUnit/UnitPreparation/**PreparationAgent** (also a Contact type) with the preparation date.

2.10. AGU: The element **ImageURI** provides a links for images in ImageType. Here we probably need a second element to accommodate strings including HTML, Javascript and the like where providers accommodate instructions how to access the images.
WGB: Added element **ImageURIString.**

2.11. MDO: I would like to see a globaly unique identifier for an ABCD unit record, which is independent from the existing 3 parted key made of institution/collection/cataloguenumber.
It is meant for globaly referencing exactly this unit record, which might be different from other unit records held elsewhere, but still talking about the same specimen with the same inst./col./catalogue number. The 3 parted key would still be very important to locate the physical specimen. GBIF is currently examining the use of LSID (Life Science Identifiers) for this or other central registry mechanism to guarantee global uniqueness. As this GUID format is not known yet, the attribute would have to be optional and of type xs:string.
WGB: Added an optional element **UnitGUID** typed String.

** 2.12. MMA (translated): A unit can be specified as to sex [in UnitCollectionDomain/ZoologicalUnit] and number of individuals [in UnitMeasurements]. However, especially in zoology a single unit is often comprised of male and female individuals or mixed with larval stages. Mapping this to ABCD would inflate the number units. I suggest to add at least attributes for number of male, number of female, and number of workers (for ants etc.).
WGB: Introducing a controlled vocabulary or even element/attribute names seems to be very specific. This is a more general problem, for which originally the MeasurementsType was conceived. However, using it now would mean that we end up with two places for data. The issue is under discussion.

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.

3.14. JKE: You explained that relationships between different identifications of a single unit (e.g. host/parasite) can be expressed by two identifications with different roles. How is that role defined?
WGB: The attributes and elements of the Element UnitDataType/Identifications/Identification further describe the role of an identification in the unit-context:
Attribute PreferredIdentificationFlag to designate current identification. In cases where more than one name applies to a single unit, several identifications should be formed and marked as preferred. Attribute NonFlag to designate negative identifications. Element Role to designate the role of the identifcation result (e.g. substrate/isolate, host/parasite, etc.). This should not affect the Names&Concepts standard.

3.15. JKE: Why are TaxonIdentied and MaterialIdentified not alternatives?
WGB: They are now (post v. 1.30) - under a new element IdentificationResult. This will allow further extension of the identification type to cover other areas (e.g. minerals).

3.16. JKE: What is the purpose of the HigherTaxa/HigherTaxon element in the TaxonIdentifiedType?
WGB: It represents a classification of the identification result, not an identification result in itself. It is used (and demanded) by collection holders as a means to include their classification into the result. It is also useful for information access as long there are no complete and effective taxonomic thesauri available for query expansion.

3.17. JKE: Are the elements ScientificNameString and AuthorString not just duplications of elements under NameAtomised and FullScientificNameString? Is this third level of

representation of names really necessary?
(DHO expresses similar opinion, see under 3.25 below)
WGB: This was maintained only for compatibility reasons (Darwin Core). Both elements have been removed for ABCD v. 1.43.

3.18. JKE: What is the role of the elements NameAddendum and IdentificationQualifier under ScientificName in the TaxonIdentifiedType?
WGB: They represent additions to the scientific name in the strict sense - used to express an insecurity of the identification (IdentificationQualifier) or a specification of the concept (NameAddendum). They are only to be used when a scientific name is the result of the identification. From v. 1.40 on, the scientific name was turned into the ScientificNameIdentifiedType, which in turn is an extension (for ABCD) of the ScientificNameType, which consists only of the name elements adhering strictly to the codes of nomenclature, namely: NameAuthorYearString (now renamed FullScientificNameString), ScientificNameString, AuthorString, and NameAtomized (entire container). The extension ScientificNameIdentifiedType includes the non-code items NameAddendum and IdentificationQualifier.

3.19. JKE: If the result of identification is the name of a higher taxon, where is the scientific name to be placed?
WGB: Originally, we had it covered by HigherTaxon, but this is now only to be used for classification purposes (see above). Neither the description of the FullScientificNameString nor the atomised name structure currently (v. 1.30) accommodates a suprageneric monomial as the result of an identification. I have changed the annotation of the FullScientificNameString and the Genus elements (now: GenusOrMonomial) in Botanical and Zoological structures accordingly (where the Codes define higher taxa - I have to look this up for bacteria, as far as I remember they don't exist in Viruses).

3.20. PME: The possibilities suggested by WGB for the concept of “isolated from” (see OECD minimum dataset .. and ABCD 1.30, 28. Jan. 2004) are interesting but the information is impossible to split in these different categories for the existing databases, where names and materials are usually mentioned in the same ‘substrate’ field.
WGB: A search for the host or substrate organisms together with other records of that organism would of course be very interesting (e.g. “Are there cultures of microorganisms from Citrus?”). For the time being, I suggest to put all substrate records into Identification\IdentificationResult\MaterialIdentified.

3.21. MJA: I can't see an intuitive way to identify the typified name and the name stored under
WGB: The typified name is in ABCD stored together with the information on the verification of the type status etc. under NomenclaturalTypeDesignations. We have long pondered if to treat this as just another identification, but there are important differences: the only possible result is a scientific name in the strict sense (see v. 1.41 complex type) and there are several addtional data items which would extend and confuse the normal unit Identification construct. Furthermore, if treated as a simple identification and flagged, it would become ambiguous once several designations are made for a single specimen. The name should thus be entered in the NomenclaturalTypeDesignations area when properly veryfied; nevertheless, it can also be cited as a "normal" identification. In

contrast, the NameStoredUnder is just another type of taxonomic identification and can thus be covered by an attribute or an extension of the identification type. In v. 1.4 now a flag in the IdentificationType is used to store it alongside with other flags.

3.22. MJA: there is no field for the species author [for names of infraspecific rank]
WGB: This is part of the systematic hierarchy and not required by the code. We are open for discussion on this point, but including the systematics of a name in a schema focussing on units is - if at all deemed necessary - a transient measure, since this function should eventually be replaced by a taxonomic system. However, the same argument holds true for the HigherTaxon hierarchy included - it is just there to facilitate searches for the time being.

3.23. YDJ: I found non-atomised types for authorship and year citation for the 'NameZoologicalType'. An explicit choice?
WGB: Yes. After lengthy discussions we decided not to further atomize author citations, neither into name(s) + year in zoology, nor into name(s) and ex-author name(s) in botany. This is a pragmatic decision; in the zoological case, there is good reason to separate the year, but this would mean introducing year and authors without year as two new elements.

3.24. DHO: For consistency can't Genus become GenusOrMonomial for NameBacterialType and NameViralType? In each case it may be a family or higher.
WGB: For bacteria this is true and the name has been changed. For Viruses the nomenclatural status of monomials is less clear, but common usage include "family names", so this was changed, too.

3.25. DHO: Regarding 3.17, I am increasingly concerned to avoid unnecessary proliferation of different formats. In the GBIF Portal I have an enormous amount of code dedicated to attempting to parse scientific names from the Darwin Core records we are retrieving. The number of ways that data get mapped from databases to the different fields is a nightmare with even the relatively few, well-explained elements in that schema. Some problems are formatting problems (entirely upper case, entirely lower case). Others involve an inability to atomise data to the level required. The ScientificName element in that schema is intended to hold just the genus and epithets, with the ScientificNameAuthor holding just the author citation. I have had to work with providers providing a combined string with all of these elements in the ScientificName field, others who provide everything in one field and parts in the other, and yet others who put everything in both fields. At the same time I have all of the Darwin Core atomised fields to consider. This means that automated parsing of names is fantastically complex (something which does not show up so much if the data are simply formatted and displayed to users as a set of elements). With even more options to consider with ABCD it could become almost impossible to ensure that different tools interpret the names in a consistent way. From the standpoint of software processing I would prefer to see all names collapsed into a much more minimal set of elements. (What is a provider supposed to do if it currently holds a database with records spanning different nomenclatural codes? It could be very hard to generate appropriate NameXxxType elements for everything.) I would strongly recommend moving towards a more simple model, such as:

TaxonIdentifiedType	HigherTaxa	HigherTaxon*	
	ScientificName	FullScientificName	
		Name Addendum	
		NameAtomized	GenusOrMonomial
			Subgenus
			FirstEpithet
			SecondEpithet
			Rank
			HybridFlag
			AuthorString
	CombinationAuthorString		
	CultivatedPlantNameElements		
			Breed
			NamedIndividual
		IdentificationQualifier	
	InformalNameString		
	NameComments		

WGB: The first level under TaxonIdentifiedType is identical under ABCD, only that ScientificName and InformalNameString are explicit choices. At TDWG in Lissabon, there was some discussion about this point, and it is clear that we always have some garbage in the data, but this should be resolved by either placing all as informal names or to or live with the dirty FullScientificName data (and filter them out at the portal's side in the case of name queries).
For the ScientificName in your schema (=ScientificNameIdentified in ABCD) we have also the same elements, only that the identification-specific items are put together as a type extension of the scientific name itself.
The differences is in how to treat atomised names. Your solution strongly resembles the one in an earlier version of the schema. After some discussion, we thought that it makes sense to separate the names into the different codes, for several reasons, among them:
Leaving the communities in charge or "their" name structures
Data integrity rules differ among the Codes and may change differently in the future
Comparison against standard nomenclators is facilitated (e.g. Bacteria!)
Element names can be named according to community usage (e.g. what's the author string?
All this and some more I don't recall right now led to the present structure. However, we anyway have to come back to this question once we have the Names standard defined.

3.26. MDO: Why do the names of the concepts FullScientificNameString, ScientificNameString, and AuthorString contain the term "String"? This is not the case in other elements. Should this be omitted?
WGB: I stand to be converted, but I think this is the only case in the schema where elements actually represent a concatenated string of other elements (those in the NameAtomised section).

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.

4.06. MDO: All agents should have the possibility to provide a LogoURL to be used in interfaces. Currently this is only possible as a supplier.
WGB: Done.

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?

5.04. AHA: GatheringAgent (/DataSets/DataSet/Units/Unit/Gathering/GatheringAgents/GatheringAgent): basically the same question as the one about NamedArea (6.12.: problem of documenting the sequence of element repetitions) applies here as well: the sorting order seems to depend on the order in which elements are delivered by the provider database which is not very reliable, except for the first collector of a team (PrimaryCollectorFlag). Also here I would like to find something equivalent to a sequence number (attribute?). The fast alternative is using concatenated output in GatheringAgentsText instead, but I assume the aim is to provide the atomised data where they exist.
WGB: The same answer as under 6.12 applies: the sequence of GatheringAgent tagged texts are the correct sequence in the XML document, the query has to ensure to include any provision made for sequencing in the source database, the wrapper has to write that correctly into the document. An interesting integrity check would be to see if the primary collector is also the first.

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).

6.07. JWI: .. I was searching in the ABCD schema for elements to capture the paleontological concepts of Period and Epoch and I was unsuccessful. Do they exist? If so, where are they to be found? If they do not exist, should they? They are strange concepts to categorize since they are kin to collecting events, but not in the same way as for non-paleo disciplines.
WGB: Period and Epoch are indeed site descriptors within the context of the collection event. They are covered by the ChronostratigraphicTerm Element in the StratigraphyType (now part of the ExtensionPalaeontological.xsd sub-schema). In ABCD, they belong to the Stratigraphy Element of the GatheringSite.
ARI: Period and Epoch are "used and abused" chronostratigraphic terms. General acceptance is that the terms "System" and "Period" are interchangeable, and "Series" and "Epoch" are interchangeable. Caution is required when looking at actual usage. The International Commission on Stratigraphy chart is a good starting point: http://www.eas.purdue.edu/chronos/Divisons_GeolTimeUSGS.pdf but you'll find many variations (for instance the chart the British Geological Survey uses).

**6.08: SRO: Form my point of view the wording of following elements is a bit confusing: LocalityText, AreaDetail, Site-Coordinate. Following the comments (January 2004) I understand that LocalityText is the original text on the label and AreaDetail is for additional information (not given on the label, like fieldnotes?). I would call the descriptive data SiteText according to Site-coordinate instead of LocalityText. SiteDetail is for descriptive data like field notes.
WGB: LocalityText is the original label (or field notes / original data entry) text. The following elements are atomised elements within that text, namely areas and further descriptive text (given on the label, i.e. not like fieldnotes). Indirectly you touch upon an important issue – how to distinguish derived data from original data. Up to now, this is done only in the country type, where there is a single element for a derived country name (e.g. the English translation). We will further discuss this issue.

**6.09. SRO: LongitudeDecimal, LatitudeDecimal. Are the elements Degree/Minutes/Seconds removed? Of course the decimal degrees are easier to handle, but many providers do only have DMS and are not willing / able to calculate decimal degrees. (In addition errors are more obvious and easier to detect looking at DMS).
WGB: **[under discussion in BGBM networking group]

**6.10. AKR: Lat/Lon can be represented using an ISO Standard (see http://www.ftp.uni-erlangen.de/pub/doc/ISO/english/ISO-6709-summary):
ISO 6709:1983 "Standard representation of latitude, longitude and altitude for geographic point locations" a format designed for usage in human readable compact file formats, protocols, etc. The standard allows both a minute/second representation as well as a decimal fraction representation.
Latitude can be represented as
DD.DD degrees and decimal degrees
DDMM.MMM degrees, minutes and decimal minutes
DDMMSS.SS degrees, minutes, seconds and decimal seconds;
prefix with + north of and on equator, and with - south of equator.
Likewise, longitude can be represented as
DDD.DD degrees and decimal degrees
DDDMM.MMM degrees, minutes, and decimal minutes
DDDMMSS.SS degrees, minutes, seconds, and decimal seconds;
prefix with + east of and on prime meridian (Greenwich), and with -west of Greenwich up to the 180th meridian.
Leading zeros are required for latitude and longitude.
Optionally, append altitude in meters (prefix with + above and on the geodetic reference datum and with - below it).
If a termination character is needed in the format, / is recommended.
Examples:
+40-075/
+401213.1-0750015.1+2.79/
+40.20361-075.00417/
**WGB: **[under discussion in BGBM networking group]

6.11. BRI: I'm particularly interested in how I might go about transferring numeric geographic data, such as a geocode taken using a GPS device, and the datum that GPS was set to when the value was collected?
WGB: I think this is covered by elements under SiteCoordinates in the GatheringType, as long as your “geocode” represents coordinates. The usage of the GPS itself is covered by CoordinateMethod, the SpatialDatum and the coordinates themselves unter CoordinatesLatLon.

6.12. AHA: NamedArea (/DataSets/DataSet/Units/Unit/Gathering/GatheringSite/NamedAreas/NamedArea): I cannot quite reproduce how hierarchical structuring/sorting of this repeatable group of elements works. If I understand correctly, NamedAreaClass is to carry categories (like county, region, city), and its sibling NamedAreaName the place name. Unless there is some convention for the value of NamedAreaClass, however, from user interface perspective I cannot see how to determine the sequence in which to display these elements: this will only depend on the sequence they are delivered by the provider database, but no sorting is possible to determine that "county" always comes before "region" (especially if there is no controlled vocabulary). Should sequence information be given in an attribute?

WGB: Controlled vocabularies would not work here, because area categories vary widely and their hierarchy may even be reversed in a single country (I think I remember that this may be the case in Russia?). In the XML document itself the sequence is given by the sequence of named area tags. The query/wrapper must make sure that, if a sequence is given in the database or by a convention on the provider's side, this is correctly translated into the sequence in the XML document.

6.13. AHA: SiteFeature/Domain (Unit/Gathering/GatheringSite/SiteFeatures/SiteFeature/Domain): Is there any standard list of terms for this (planned)?
CCO [Prompted by WGB]: The site feature domain serves to distinguish items such as earth science features (e.g. fossiliferous horizons) from biological features (e.g. a veteran tree or population of bats) etc. This is useful for sorting and searching and delivering appropriate term lists. I have a provisional list of domains in the thesaurus but it is likely to change with use (The thesaurus is arranged by Subject and domain as a means of distinguishing groups of term lists). It is possible that a field record or a specimen be directly linked to a site feature and this is the reason for its presence in the schema (In the larger data model features can have data of their own including management actions, threats, damage etc.). As far as immediate use of the schema goes, I think it unlikely that many existing data sets are arranged in such a way that this element would be used - however, the extended version of Recorder software (currently in beta test) does use domains and location features. The scope of the schema has been an issue from the start of the project, clearly, working towards a highly modular form that has a limited core to meet 90% of normal needs with extensions that meet more specialised requirements is the best answer - In answer to your question - it probably won't be missed from the core but shouldn't be forgotten completely.
WGB: I removed the entire domain for the time being because it is currently not used and could be confused with other elements.

**6.14. AKR: The marine XML schema documentation provides another example for co-ordinate data that may be re-used for ABCD. [See Appendix 1].
WGB: Under discussion.

**6.15. TC: The format of geographic coordinates in ABCD schema was examined. The latitude and longitude attributes are defined as String type in the Schema and it's probably an error that should be checked and numeric type applied instead of String.
WGB: Done. Should we include type for range of allowed values? What about Datum?

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.

7.03. AKI: the following rule should read '....between 01 and 24'

<Rule xml:lang="en">The hour is expressed as a 2-digit value, left zero padded if necessary, ranging between 01 and 31.</Rule>

WGB: Corrected.

7.04. MDO: the ABCD DateTime type should include examples and a detailed explanation of the ISO format:

Year: YYYY (eg 1997)

Year and month: YYYY-MM (eg 1997-07)

Complete date: YYYY-MM-DD (eg 1997-07-16)

Month and day only: --MM-DD (eg --07-16)

Day only: ---DD (eg ---16)

Complete date plus hours and minutes:

YYYY-MM-DDThh:mmTZD (eg 1997-07-16T19:20+01:00)

Complete date plus hours, minutes and seconds:

YYYY-MM-DDThh:mm:ssTZD (eg 1997-07-16T19:20:30+01:00)

Complete date plus hours, minutes, seconds and a decimal fraction of a second

YYYY-MM-DDThh:mm:ss.sTZD (eg 1997-07-16T19:20:30.45+01:00)

where:

YYYY = four-digit year

MM = two-digit month (01=January, etc.)

DD = two-digit day of month (01 through 31)

hh = two digits of hour (00 through 23) (am/pm NOT allowed)

mm = two digits of minute (00 through 59)

ss = two digits of second (00 through 59)

s = one or more digits representing a decimal fraction of a second

TZD = time zone designator (Z or +hh:mm or -hh:mm)

WGB: examples and comments were included in the AppInfo.

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.

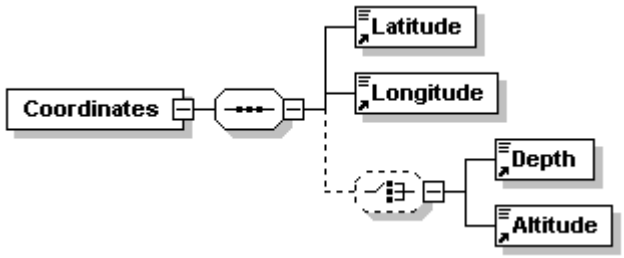
**8.05. MDO: A controlled vocabulary should be provided for Measurement units
WGB: This should be discussed – the measurement and fact types will be revised.

App. 1: MARINE XML SCHEMA DOCUMENTATION

Source: http://www.aodc.gov.au/products/prod/documentation/marine_xml_schema.html#GeoPoint

Coordinates

The *Coordinates* element is used to define the coordinates of a single point in space. It was decided that the coordinates element would contain child elements *Latitude* and *Longitude* for describing marine spatial data. There is also an option to include a *Depth* or *Altitude* child element if required.



Child Elements:

Element	Type	Number	Description
Latitude	Required	1	Used to specify the latitude.
Longitude	Required	1	Used to specify the longitude.
Depth	Optional / Choice of	0 or 1	Used to specify the depth.
Altitude		0 or 1	Used to specify the altitude.

Attributes:

Name	Type	Use	Description
datum	String	Required	Used to define the datum used for the latitude and longitude measurements.

Latitude

This element encases the decimal value of the latitude of the record. The precision used is up to the user. It is suggested to use a value to at least six significant figures.

Negative latitudes southern hemisphere.
Positive latitudes northern hemisphere.

Longitude

This element encases the decimal value of the longitude of the record. The precision used is up to the user. It is suggested to use a value to at least six significant figures.

Negative longitudes western hemisphere.
Positive longitudes eastern hemisphere.

Depth

This is an optional element. If depth is required as part of the spatial coordinates for the record then this element should be used. The depth element encases the decimal value of the depth. Attributes are used to assign datum and unit information. The *Depth* and *Altitude* elements cannot both be specified.

Attributes:

Name	Type	Use	Description
datum	String	Required	Used to specify the datum used to determine the depth.
units	String	Required	Used to specify the units of the depth measurement. Suggested unit of measurement is "Metres".

Altitude

This is an optional element. If altitude is required as part of the spatial coordinates for the record then this element should be used. The altitude element encases the decimal value of the altitude. Attributes are used to assign datum and unit information. The *Depth* and *Altitude* elements cannot both be specified.

Attributes:

Name	Type	Use	Description
datum	String	Required	Used to specify the datum used to determine the depth.
units	String	Required	Used to specify the units of the depth measurement. Suggested unit of measurement is "Metres".

Figure 1