

Minutes of the Annual Meeting of the Taxonomic Databases Working Group Christchurch, New Zealand, 11 – 17 October 2004

Attendance

The 2004 Annual Meeting in New Zealand, kindly hosted by Landcare Research (New Zealand), was attended by 115 delegates from 86 organisations in 26 countries.

Meeting Summary

A list of sessions and presentations is at: http://www.tdwg.org/2004meet/TDWG_2004_Papers.htm

The very full agenda was arranged thus:

- Day 1 TDWG Mission, Structure, & Process (for new delegates)
Architecture for Biodiversity Data and Services, including Globally Unique Identifiers (GUID) in Taxonomic Databases (LSID – Life Science Identifiers)
Access to Biological Collections Data (ABCD) - ABCD Schema
- Day 2 Darwin Core Schema
Distributed Query Protocol - The integration of DiGIR and BioCAsE
Structure of Descriptive Data
- Day 3 Taxonomic Names & Concepts
Unified Biodiversity Information Framework (the “glue” binding the various elements of biodiversity data & protocols), including presentations by staff of GBIF
- Day 4/5 TDWG Business
Presentation of general papers
Spatial Data, Observational Data and Imaging
Proposals for new Subgroups and interest groups
Presentation of a revised framework for coordinated standards development
- Day 6/7 Subgroup working sessions

The early part of the meeting was dominated by the universally recognised need for the application of Globally Unique Identifiers (GUID) and the directly applicable Life Science Identifiers (LSID). This relates directly to the framework for taxonomic names and concepts being developed as a standard by the appropriate TDWG subgroup in association with members of the SEEK Group. Following on from this was discussion on the bringing together of the ABCD/BioCAsE and Darwin Core/DiGIR strands for standardised data and delivery. This is deemed to be of crucial importance, as is the outline of future developments introduced during the Unified Biodiversity Information Framework (UBIF) session. The session on taxonomic names & concepts produced lively discussion, ending in a remarkable level of agreement and a timetable for the submission of a draft standard. All active themes, including Structured Descriptive Data and Spatial Data, are now working towards a common framework.

Staff of GBIF gave presentations describing the building of a Global Network using TDWG standards, providing a comprehensive overview of the way that the GBIF will work in the future to include TDWG-originated developments. GBIF reiterated the crucial importance of emerging TDWG standards to its ability to collate and present data from disparate source to the international community. The meeting was reminded that the GBIF has severely limited core funds and that the subscribing members have a duty to support national and joint initiatives directly.

Towards the end of the plenary sessions emphasis was placed on spatial/geographic data and joint progress with work in the external georeferencing community towards common standards and

interoperability. This included a presentation on the work of OBIS and notification of the Ocean Biodiversity Informatics conference due in early December. Finally, the broad new topics of managing observational data and digital imaging were introduced as a new theme.

There was general agreement that this had been one of the most successful TDWG meetings ever, and probably the most exciting, at a high technical level. The standard of presentations and discussion was high. There is now a firm framework for the establishment of a Steering Committee to coordinate the expanded subgroups and efforts to bring the various strands of TDWG activities together. A "charter" for standards development will be posted on the TDWG web site. There will also be a new web site archive.

Following much discussion in meeting sessions and in breakout groups new Subgroups and special interest groups were provisionally established to take forward less well developed subjects:

- Natural Collections Descriptions (convener: Neil Thomson, NHM)
- Observation/Monitoring (convener: Steve Kelling?)
- Image Content
- Bibliographic References & Taxonomic Literature

Social Events

Trips were organised before the start of the Meeting to Arthur's Pass (breathtaking mountain scenery) and Kaikoura (Whale Watching).

An evening Welcome Reception was held at the Heritage Hotel. The Meeting Banquet took place in the Curators House Restaurant in the Botanic Garden. Both events were well attended and enjoyed by all.

TDWG 2005 & 2006

It is very clear that the next two years will be intense as the various themes converge and a number of draft standards are presented for adoption. Subgroups will be highly active during the year in preparation for the next Annual Meeting, likely to be held in St Petersburg (Russia) in September 2005, hosted by the Zoological Institute of the Russian Academy of Sciences. It is anticipated that the 2005 meeting will attract a large number of delegates, requiring a more formal structure and tight criteria for the acceptance of presented papers. The 2006 meeting is provisionally scheduled for October in New Orleans.

Account of Meeting Sessions

A summary of discussion at the Plenary Sessions is attached. Proceedings during individual subgroup sessions are not included.

Adrian Rissoné
TDWG Secretary

TDWG 2004 - Day-by-Day Summary of Meeting Discussion

These notes were compiled during the various meeting sessions. Apologies to anyone who feels they have been misrepresented.

Monday 11 October

Morning Session

Welcome by Stan Blum
Housekeeping by Jerry Cooper
Welcome by Andy Pearce (CEO, Landcare Research)

Stan Blum

TDWG Mission

TDWG Structure

- Membership
- Executive Committee
- Subgroups

Overview of meetings since 1985

TDWG Process

- Structure
- Mechanisms
- Subgroups – how do they work?
- Membership – different models
- Relationship with GBIF

Discussion:

Arthur Chapman asked for an outline of the future, including names and versioning of TDWG standards.

Stan Blum said he thinks a repository is the most important thing, well documented. Most standards are developed by individuals (project groups).

John Wiczorek. thinks it would be useful to spread the word about TDWG.

Ben Richardson echoed that citing Mozilla as an example.

Chris Frazier, TDWG has outgrown its name.

[Unknown speaker] – which standards are current? Which retired?

Stan Blum – TDWG is run by volunteers. Please help things along by getting involved. Use mailing lists. Contribute to archive. Better documentation and history by subgroup conveners so that people can see what has happened.

Chuck Miller – what do the by-laws say about this?

Stan Blum – nothing in this meeting will require a vote. Maybe we need to change the by-laws to allow things to happen between annual meetings. Stan Blum would like to change the Executive structure away from regional members towards a proper steering committee. Maybe subgroup conveners should be members of the steering committee. Harden up standards process to have proper way and end points.

Sally Hinchcliffe – use a newsletter (resurrect the Newsletter Editor).

Giorgos Ksouris - Overview of XML Technologies

Giorgos gave a presentation explaining XML and related technologies.

Bob Morris – don't panic! Tools and frameworks are available to generate all the nasty code.

Chuck Miller – XML header – what about the encoding? If your XML document includes non-ASCII characters the end document could be unreadable.

Jessie Kennedy – make sure defined encoding is correct.

Chuck Miller – example, Word internal encoding extended ASCII does not work with UTF-8

Gregor Hagedorn – shouldn't be a problem.

Dave Vieglas – source data document(s) must be encoded to match the header

Charles Copp – BioCASE thesaurus uses SQL-Server (source data) and Postgres (delivery) needs a translation module between.

Chuck Miller – the big issue is that there is a problem matching streams of data.

Bob Morris – if you say it's UTF-8 it must be UTF-8! It's not the fault of XML.

Dave Thau

Dave gave a talk on Life Science Identifiers – a type of Global Unique Identifier - with amusing asides on the acronym "GUID".

Do we need GUIDS?
What gets a GUID?
One GUID per item?
Centralization?

Discussion:

Stability?
Changing data – conceptual conflict with GUIDs?
Versioning?
Backwards versioning?
Caching?
Discarding of versions?
Internal structure (of versions) must be fixed
Client could choose between versions? E.g. Genbank
Degree of granularity? Depth?
Why better than other kinds of metadata?
GUIDs will develop naturally (organically)?
There are things already looking for GUID solutions – need to make some progress soon
GBIF to be holding workshops [Donald Hobern]

Afternoon Session

Architecture for Biodiversity Data & Services

Stan Blum - primer

Ways of looking at architecture:

Goals

Data Dynamics – who, what data, what services/protocols?

Scaling – how complex before buy-in attenuates?

Object Identity, Persistence & Annotation – what happens when things change?

Is an annotated object the same or different?

Systems issues – performance, interoperability, availability, recovery, planned redundancy

Miscellaneous – quality & consistency, priorities & valuation, common features of data & services, functions to support common features

Discussion:

Greg Whitbread – implication of LSIDs for protocol group?

Stan Blum – more impact for schemas. Darwin Core is a close match to the way people do business so they will continue to use existing identifiers. LSIDs will be introduced gradually.

Greg Whitbread – LSIDs will resolve to records?

Donald Hobern – wait and see what emerges over the next few days (eg. Markus & Renato). Thinks we have a springboard to move to a wide variety of object types. The main problem is knowing what we want/need to share. Danger of overlapping, ill-defined data types. Thinks we could use LSIDs extensively – could be used by BioCASE, DiGIR, OpenGIS, wherever possible. Things need to come together – schema repositories, etc.

Need to identify use cases – real situations – to define object types, etc.

Greg Whitbread – how would this work?

Donald Hobern – use LSID approach to resolve where data sources are – the discovery mechanism. Versioning could be useful.

Stan Blum – complete DiGIR record, complete ABCD record? What? The same objects can have different views in different contexts.

Gregor Hagedorn – split protocol? Abstract object returns a specific data set. Plan this into the protocol.

Dave Veiglais – LSID resolution service becomes lowest level for accessing data. So protocol should access the data through an LSID system (via metadata?). Simple data provider, indexing layer, and so on.

Bob Morris – social issue – never been a contract that a given query always returns the same thing.

Greg Whitbread – an LSID could return the same query.

Gregor Hagedorn – query returns LSIDs, use LSIDs to retrieve data. Need to split protocol to return only (say) a few fields. An LSID assumes that you get all the data for an object.

Dave Veiglais – use LSID as the access mechanism, not the filter – that is the responsibility of the data provider.

Dave Veiglais – need consistency

Greg Whitbread – the LSID assumes consistency internally. LSIDs allow layers and abstraction.

Bob Morris/Greg – discussion on internal consistency of LSIDs

Walter Berendsohn – all our data may change over time so how do we fulfil the LSID condition?

Bob Morris – our community is only interested in LSIDs at the metadata level (except for abstract objects?)

Dave Thau – LSID often are metadata and not the actual object behind them

Jessie Kennedy – change ends up in something different. Need to define what changes the original definition and what does not. There are some things that cannot change for an LSID to remain valid

Bob Morris – example of misidentification – not enough data provided to the identifier. What's the data and what's the data about?

Gregor Hagedorn – descriptions may change so the LSID needs updating – keen on versioning. The problem is the restriction on changing only a certain set of bytes in an LSID.

Sally Hinchcliffe – difference between physical objects and taxon concepts. A specimen is part of the real world.

Jessie Kennedy – why do specimens change? What people say changes but the physical thing does not change.

James Ytow / Jessie Kennedy / Chris Lyal – discussion of what constitutes a change.

Walter Berendsohn – why not use existing reference systems and choose what is most relevant?

Sally Hinchcliffe – you are still looking at the same object no matter what has been value added but the metadata does change

Rich Pyle – complication of lots being divided up, even going to different institutes so you need a new GUID

Jessie Kennedy – apply only to a single object

Discussion of what constitutes an object.

Gregor Hagedorn – versioning is the way to handle these changes. Subsequent queries must be able to return the same data, changes are via versioning. But if versioning is not maintained data may be out of date. Also, asks for a definition of metadata.

Arthur Chapman – an LSID is permanent and may extend beyond the lifetime of the object itself.

Chuck Miller – notion of uniqueness. The process is layered communication protocols to drill down to the unique ID.

Ben Richardson – what about the Director's edition of a movie?

Anna Weitzman – splitting collections happens all the time. Numbering is a problem. Gave an example of a whale, the parts of which were in many institutions, giving a false genetic bottleneck.

Greg Whitbread – we build dumb systems for smart users. For most purposes it's arbitrary.

Anna Weitzman – OK, but we need a way to track specimens

Bob Morris – LSIDs do provide a mechanism but it requires a social agreement.

Gregor Hagedorn – context and granularity influences how you use LSIDs – in some situations there may be difficulty. Need to allow for reasoning.

James Ytow – what is the distinction between identical and same?

Charles Copp – where is the reality in all this? There are databases that track changes. It's no good just inventing a new numbering system because it will not be used. Resources not there. So, anything new must be grafted on top of existing systems. So, what can we use in practice?

Jessie Kennedy – What would be the benefit of all this? The way things are at the moment there is great duplication of effort and numbering. LSIDs could be a big benefit if introduced systematically.

Walter Berendsohn – we're been discussing theoretically. We need to know how to use LSIDs practically.

Anna Weitzman – whatever we come up with must be usable in the Museum community. Otherwise they will be misused.

Rebecca Shapley – need a system of relationships between LSIDs and categories.

Bruce Stein – what are the costs? Curating LSIDs is a problem. Can't be allocated automatically. What's the minimum set of things that would provide a real benefit.

Bob Morris – supposed to be semantically opaque. So relating LSIDs is inappropriate. It's the schemas that need to be defined.

Gregor Hagedorn – agrees with Bob Morris. The problem is that we are trying to invert the sense, which is wrong. In practice duplications will occur. This has to be resolved by a different service – this will cost money and will only be done where there is a perceived benefit.

[unknown speaker] – who solves what? How do we start?

Rich Pyle – Costs on multiple schemes of UIDs but the benefit may be worth it.

Greg – starting to see a place for UIDs. There will always be ways to resolve different schemes of UIDs

Donald Hobern – against imposing a ridged scheme where it may cause problems. Can use proxies to assist.

Jessie Kennedy – multiple LSIDs are a worry. The benefit of an LSID is that you know you are taking about a unique object.

Gregor Hagedorn – same object, yes, but is it worth the cost to resolve duplication?

Jessie Kennedy – could get out of control and become meaningless

Stan Blum – need to wrap up the session. Donald Hobern has some small resources but hopes to be able to set up a serious review, probably soon (next couple of months?).

ABCD

Walter Berendsohn gave a resumé of the principles, development and future of the ABCD standard.

Discussion:

[Unknown speaker] – Is BioCASE/ ABCD compatibility in place. Walter Berendsohn said “yes” and explained certain restrictions.

Charles Copp – changing the model move moves away from the specimen view previously inherent. This shows that we can make complex types.

Walter Berendsohn – much of the voting will be on the individual element definitions. It consists of building blocks from earlier versions and provide a framework for the future, including complex structures.

Dave Veiglais – should we be considering the use of LSIDs?

Walter Berendsohn – (from Markus) every element should have a database ID. We can use IDs wherever we can define an “event”, otherwise only highly regulated things such as names can be effective.

[Unknown speaker] – how do you define geographic information?

Walter Berendsohn – when I find a collection with GML data I can make rapid progress!

Phil Dibner – I’m interesting in hearing about this [idea to talk later]

Dave Veiglais – thinks there are data sources from OBIS with GML [nobody from OBIS present]

Joanna McCaffrey – what is the relationship with Darwin Core 2?

Walter Berendsohn – extensive mapping between schemas. Problem is installed base – users think DC2 is enough. Both standards have a role at the moment. It gets complex is you want to map ABCD to DC – only available via BioCASE, not DiGIR

Stan Blum – we will talk more about DC Tuesday morning. We [DiGIR?] are interested in keeping to a smaller core – simple taxonomic descriptions and maybe simplest descriptive data. Ultimately we need to make as compatible as possible. Taking up in next 6 months or so.

Chuck Miller – is there an installer now? DiGIR has one through GBIF

Walter Berendsohn – yes, available on the web

Javier de la Torre - can install in 5 mins! Look on the BioCASE web site

Jessie Kennedy – detail of Identification event please?

Walter Berendsohn – I thought you would show it tomorrow! There are not many changes.

Bruce Stein – how robust is the current version relative to observational data types?

Walter Berendsohn – what do you mean by robust? Yes, it can handle large numbers of records. It’s just that observation data is more extreme – the principles are the same. We will have to adapt the model and the protocol as we go on.

Reed Beaman – are any collections using spatial data (not just GML)? What would be the best way to handle those data. The spatial data group will address that later in the week. GML has a place to make interoperability as good as possible.

Walter Berendsohn – there are many parts of the schema that could be improved but we need to get on with it – to mobilise the data. We don’t yet have the critical mass of data for the standard to run on its own. We will have spatial data, then it’s the portal guys who have the problem.

[Unknown speaker] – returning to the importance of observational data.

Walter Berendsohn – the area is so large that it will take time to work through it. We can’t wait for that to happen. We should use the existing schema areas for now and build on it. There are, in fact, lots of observational records on line.

Donald Hobern – yes, but it’s being treated as taxon occurrences. But, they relate to much more information we need to leverage.

Walter Berendsohn – we will continue to bridge to external efforts. What we must avoid is duplication.

Stinger Guala – the NSF has supported a lot of effort, mostly DiGIR. Now BioCASE is appearing at the same highest levels in Government. It all depends on not a merging of DC2 and BioCASE but there must be something usable. There is a lot of money potentially available and this meeting is important in finding ways. He has been able to explain the convergence to senior position in the U.S. Government so it's up to us to follow through.

Walter Berendsohn – the effort to converge is considerable.

Stan Blum – we will be discussion this tomorrow morning (Tuesday) on this reconciliation effort between protocols (DiGIR/BioCASE). It is a difficult subject and there will be issues.

Gregor Hagedorn – ABCD and DC2 are different and each has its place. DC2 is an interface to specimen data – a short version – ABCD is broader. Improvement can be achieved by e.g. using the same elements names.

Jessie Kennedy – does not really agree.

Gregor Hagedorn – Inconceivable that we only work with data from major collections. We have research collections and DC would be good for this.

Jessie Kennedy – why not use a subset of ABCD?

Dave Veiglais – most important thing is that DC data can be derived from ABCD. Then DC becomes a commonly used subset of ABCD. DC is used for small representations of data, ABCD is much more complete.

Walter Berendsohn – this was exactly the aim 3 years ago. There are areas of inconsistency and some real problems. The providers have to come around a bit as well.

Chuck Miller – your used the term protocol. But also the differences between the schemas mean that you have to use different protocols.

Stan Blum - the DC schema will work with BioCASE. The DiGIR protocol cannot handle a schema as complicated as ABCD (it's essentially flat at the root level). This will be discussed tomorrow.

Bob Morris – are we going to hear about all of this tomorrow?

Walter Berendsohn – Yes. We should be careful to look ahead. The reconciliation may result in a resolution of the object oriented vision. Bit by bit it will come together.

Stinger Guala – Biomoby and the new plant ontologies - have you been working on those ontologies? There are two groups in Europe.

Walter Berendsohn – no, not really.

Jessie Kennedy – the plant otology is more to do with SDD

Stinger Guala – I would like to see SDD adopt ontology

Jessie Kennedy – yes, we have looked with relation to SDD.

[more discussion between Stinger Guala & Jessie Kennedy]

Walter Berendsohn – are they using anything specific. Haven't seen anything that overlaps with ABCD.

Donald Hobern – where the reconciled protocol is going should allow us to automate the interoperability.

Tuesday 12 October

Morning Session – Chair: Stan Blum

TDWG Business

This must be Stan's last year as Chair (the TDWG Constitution limits the post to 3 years). Nominations for all Executive posts are to be given to Jim Beach (Nominations Officer for this meeting) by coffee break on Wednesday 13 October.

Darwin Core 2

Stan Blum introduced Darwin Core, giving a brief history, how things have changed and where DC is going, using the DC site at <http://darwincore.calacademy.org/>

Stan also explained changes in DiGIR/Darwin Core to decouple the schema/protocol.

He explained the Darwin Core 2 Element Definitions (using the HTML version on the site).

There was discussion on "internationalisation" and date/time standardisation for indexing. The key issue on date/time is the inclusion of the time zone. If a time zone is not included it isn't an XML date! Walter Berendsohn said that the critical date was the date embedded in the data record/set. If the date is not the same as you have in your index then you need to regenerate the index. Bob Morris said there is nothing in any of the relevant documentation to govern Walter's ideas. There was discussion on the use of update flags. Donald Hobern thinks the provider as well as the portal has a responsibility to handle date indexing. Gregor Hagedorn and Dave Veiglais emphasised that a proper XML and ISO date should be all that is required from the provider.

Taxonomic Names: discussion about the "ScientificName" field being mandatory. Walter Berendsohn says this cannot always be provided (e.g. bag of sample). The general consensus was that just a simple term (e.g. unidentified) could be used. But Donald Hobern disagreed because a simple parser might allocate this term to (say) Genus. Dave Veiglais asked why this name was necessary. Because some databases do not have parsed out, atomised taxonomic names. Neil Thomson suggested using the specific term "Collection". Charles Copp said that "not determined" was better. The real issue is requiring use of a fixed list of allowable terms. Lynn Kutner said that null entries are actually ambiguous. She also questioned using this field to hold concatenated atomised data because of the provider overhead of generating the data. Bruce Stein argued that this should be a required field so as to have one place where you can get down to the lowest level of taxonomic information available. Arthur Chapman agreed but said that even the highest level term "biota" was valid and useful. Rebecca Shapley asked what should be done about data that slip between the various ranks of the DC. Just concatenate data into the HigherTaxon field? Stan Blum said he thought that is appropriate. He said it will be a question of evolving practice. DC is essentially a hack – a compromise between competing objectives. Charles Hussey commented on the HigherTaxon field that there was a potential for great value but also for great confusion. Is there a place for an "InformalTerm" field (with restrictions) where strict higher taxon data are inappropriate. Charles Copp suggested using an intermediate thesaurus. Stan Blum was concerned that a name server could work but there is no globally populated name service. Charles Copp, did not see that as a problem for higher taxa for much longer. Jessie Kennedy said that there does not seem to be a way of recording taxon concepts. This reminded Stan Blum that there is a requirement to record the "authority" for a determination. Does this satisfy Jessie Kennedy's needs? Jessie Kennedy said that an authority is not strictly necessary – it should be embedded in the concept. Stan Blum said that this area would be advised by discussion in the Taxonomic Names sessions. Stan Blum's final comment was the he thought the consensus was the field should not be mandatory. Dave Veiglais said that the standard could be more robust by simply requiring the field to have the term "biota" if nothing else is available. Chuck Miller

agreed that a controlled vocabulary is the best approach. Sally Hinchcliffe announced that there is a genus name "Biota"! Dave Veiglais said surely we can distinguish between "null" and "zero length string", the later being that the value is known and the value is nothing. Crispen Wilson warned that there is a serious danger in using nulls in that changing a value elsewhere in the record may compromise the null value. This is provider dependent. There was further discussion which did not expand any point significantly. Stan Blum announced that TDWG would be convening subgroups to address these issues. Lawrence Way said there is a requirement for a field or fields to relate to external lists of managed names (which are already emerging) – e.g. the BioCASE thesaurus.

Geographic data: Stan Blum said that Continent is now separated from Water Body. Also, ranges for elevation and depth. There will be additional field in the Curation extension schema to handle versions of georeference data. Stan Blum asked is georeference was commonly included in Culture collections. Patricia Mergen and Stinger Guala gave examples. Larry Speers said it's very variable but becoming more standardised. Gregor Hagedorn commented on the fields required for rectangles. He suggested more data typing. There is much more in ABCD and SDD. It gives the benefit of hierarchy without compromising the flat structure. Stan Blum agreed in principle with reservations about structure. Chuck Miller referred to the existing TDWG Geography standard. Can this be reconciled with DC2? Stan Blum said it has not been attempted. There are recommendations rather than requirements. The TDWG standard had originated in the botanical world and may not be appropriate in the zoological world. Chuck Miller said that he used the standard codes and there needs to be some place to populate DC2. Alan Paton asked about derived data. Should not also the original data be included? Crispen Wilson asked about the relationship between the height of the ground and the heath in the canopy. Rich Pyle and William Ulate supported Crispen's way of doing things. Arthur Chapman commented on the need for real precision in coordinate data – i.e. a requirement to not use false precision. Bruce Stein said the focus on decimal lat/long implied point data rather than the polygon approach used in other fora. We need to handle polygon data. Uncertainty values are a start but not enough. Dave Veiglais said that some software have inbuilt mechanisms for spatial data (using openGIS). This should be kept in mind for the future as these products develop. There was further discussion on point versus shape georeferencing.

Date ranges: Stan Blum didn't think that it is necessary to get too involved in date ranges. Adrian Rissone disagreed because his experience was that date ranges are used extensively. Lawrence Way supported Adrian's view and suggested a survey of existing data sets to check on the significance. **Action Point:** Adrian and Lawrence will coordinate a survey via the TDWG listserv. There is a need for a field to record the duration of an event. [Unknown speaker] said duration implies another level of range which needs separate categorisation. There was another short discussion on date/time formats and time zones, and whether the DC should be expected to handle complex date information. There was the suggestion of adding a field for verbatim date (moving CollectingDate from the Curation extension to the core). Arthur Chapman said this was covered in Dublin Core. He thinks the duration is very important. Donald Hobern thinks this level of complexity is better handled by schema extensions rather than overload the core. He does not like the idea of verbatim elements in the core. There are added-value data which are much better managed in extensions. Donald Hobern questioned whether ranges are better handled with a point date and precision. Sally Hinchcliffe said that the absence of date ranges introduces a strong risk of missed records. Gregor Hagedorn does not think that including verbatim data overloads the core but is very useful from the user point of view. Sally Hinchcliffe responded that verbatim is what was on the label and the only way of handling this is to have date range. Donald Hobern said it's important but needs to be in a different place.

Containing Element: needs wider discussion outside of this meeting.

Stan Blum summarised by encouraging people to look at the DC web site and to make it known that they are involved in particular intentions and aspects.

John Wiczorek point out that we need to stabilise DC2. Bruce Stein notes that there is a move toward community-specific extensions. How will this process work? How will we avoid duplication? At the moment there are only a small number of extensions.

Lawrence Way urged that existing definitions from ABCD and elsewhere be used wherever possible.

Integration of DiGIR and BioCASE

Renato Giovanni and Markus Döring gave an outline of the DiGIR and BioCASE protocols, the relationships between them, and detailed description of a project to integrate the protocols. This was followed by a short period of discussion. Donald Hobern explained what the GBIF can (and will) do to facilitate the process of integration. <http://ww3.bgbm.org/protocolwiki>

Afternoon Session

SDD

Wednesday 13 October

Morning Session – Chair Jessie Kennedy

Taxonomic Names & Concepts Taxonomic Concept Schema (TCS)

Jessie Kennedy gave an overview of the work since the last TDWG meeting in Oeiras in 2003, including an overview of the project and the rationale. Her presentation gave a detailed explanation of what a Taxonomic Name is, what a Taxonomic Concept is and how they relate. The concept must allow users to record the information they actually have (Definition vs Usage).

Discussion:

James Ytow – can technology map definition and usage at the database level? Jessie Kennedy replied that the crucial issue is that a definition is a fixed event. Anything modifying the definition de facto generates a new definition.

Gregor Hagedorn – my work consists of check lists, etc., from a variety of sources, brought together and cross-checked. Part of the problem is too little effort to update and revise taxonomy. Jessie Kennedy replied that a true definition must be published in some way.

Rich Pyle said that variations should be allowed in the real world for practical usage. Jessie Kennedy said the act of definition is intentional. Rich said that unintentional definitions happen. Jessie Kennedy said this is still useful but it must be recorded somehow.

Nico Franz – not sure Rich is right. Recording every single uttering of a name just confuses. There must be a cultural understanding of citing versus publishing. – whether you are creating or borrowing.

Bob Morris – concerned that the distinction satisfies taxonomists but not the end users of biodiversity data. Taxonomists insist on accuracy of concept. The average user just needs “good enough”. Jessie Kennedy replied that the user has to decide the value of what they are using – is it good enough for me?

Robert Kukla explained the TCS schema and the use of elements/attributes in different scenarios.

Discussion:

Stan Blum asked whether the Concept Type should be mandatory. Robert said it is very valuable.

Rich asked about Nomenclatural Concept Types. Using a name is one, but is not an Original or Revision Type.

Gregor Hagedorn – nomenclatural databases surely imply use of the Original Type? This is OK but something using the nomenclatural database is duplicating? This can be avoided by using GUIDs. Jessie Kennedy - This issue is of great significance to aggregators of data (e.g. IT IS).

Barry Conn – disagrees that a Name is not a Concept. The schema is fine but some people will not understand the concept and will misapply. Many original concepts are in fact described in revisions. The default is that the author believes in the original concept.

Walter Berendsohn – We must discuss uses of the standard and whether they are appropriate. Congruence is the most important from the point of view of usage. We can use this standard to inform people outside of taxonomy.

Arthur Chapman – what is meant by a name changes over time. Splitting of taxa. Asked Jessie Kennedy if she intends unpublished names to be included. These are increasingly used (e.g. in reports). Jessie Kennedy says no reason why not. This way we can keep track of developing concepts before they get published.

Charles Copp – worried about Relationships. What do ToTaxonConcept and FromTaxonConcept mean in practice? E.g. where do Fishbase numbers fit? Robert – if there is a 1-1 relationship then you are encouraged to use the IDs. Jessie Kennedy – at the top level until GUIDs are resolvable.

Robert Kukla gave another presentation on Interpreting Taxonomic Databases and Mapping their Structures to the TCS.

Chris Lyal and Anna Weitzman gave a presentation on Taxon names and concepts: building a strong foundation for biodiversity information

Discussion:

Jessie Kennedy – people use names as if they mean something. She stressed the concept is the name *plus* the definition (e.g. determination in a publication).

Gregor Hagedorn – I want to be able to reuse other people's work without compromising the concept. The name element of the object has changed during revision but the object itself remains the same. Anna Weitzman – a name is just a label, taxonomy and nomenclature are separate things. We have to cater for the widest of interpretations.

Walter Berendsohn – We have implemented the TCS in the Berlin database. The mapping was not quite correct. We have traced the changes and they do not amount to a concept change. He does not think it's relevant for the transfer standard.

Anna Weitzman- we must concentrate on that minimal set that GBIF needs to get off the ground, but keep things in mind to see where we can go forward.

Donald Hobern gave a presentation on Access protocols for Taxonomic Names Concepts and Data.

Ed Donovan gave a presentation on The Species 2000 standard data set, common data model and transfer protocols.

Nozomi (James Ytow) Ytow gave a presentation on Name Usage as Principal Data Vehicle for Name/Taxon Data Exchange

Aimee Stewart gave a presentation on the SEEK Project Taxonomic Object Service: Making Sense of Names and Concepts

Dave Veiglais said that there is a lot of information available on the SEEK web site.

Jerry Cooper gave an unscheduled short talk on the Linnean Core (as it had cropped up in discussion several times).

Jessie Kennedy summarised the morning's presentations and discussions, and worked through some of the points and issues that had been raised. She returned to her list of outstanding issues, pointing out that some things had not been covered fully during presentations but would be during the

subgroup session on Saturday (16 October). Jessie Kennedy considered James Ytow's argument about usage versus concept, wondering what people thought about the relative values, and the relevance of the Linnean Core.

Anna Weitzman emphasised that we need to look carefully at name structure. But she asked Donald Hobern what he needed immediately so that GBIF can go forward.

Chuck Miller said that this is all for the benefit of someone – who? – and it must be relevant to a wide community. Which end of the spectrum presented should people go with? Mostly, people go for the middle. We should decide the best way to go so as to prevent this.

Jessie Kennedy said that they had tried to avoid the common denominator approach, rather people use the bits that are relevant to them. This is a transfer mechanism to be interpreted by the users. Anna Weitzman agreed but we must get the detail we need to build the best offering possible.

Chuck Miller said there is a lesson in what Species 2000 has done because they anticipated several audiences.

Afternoon Session – Chair Stan Blum

UBIF Session

Gregor Hagedorn gave an introduction to the need for a UBIF, detailing the proposals for the integration of SDD and ABCD. He listed the topics to be covered – Type Library, Top-Level Structure, External data interface, Basic text formatting conventions.

Donald Hobern gave presentations on (a) DiGIR and Documents (and other developments – mostly other developments!). [note change from published title!] and (b) Building a Global Network using TDWG Standards. These two presentations provide a comprehensive overview of the way that the GBIF will work in the future to include TDWG-originated developments.

Discussion:

Bob Morris – thinks the processes described will memorialise the tight relationship between protocols and concepts. Donald Hobern says he thinks exactly the opposite. Nothing he might regret is being set in concrete.

Gregor Hagedorn gave a presentation on: Proxy data objects provide optional linking to external objects and small modular data interfaces from which DarwinCore and LinneanCore- like protocol interfaces can be constructed.

Discussion:

What do you want? This is a set of interfaces, without the methods. The developer is free to propose methods to become standards. The ideal interface is clearly some XSLT script to reduce the output from DC or ABCD. However, this is not a requirement.

Jessie Kennedy – in your taxonomic example you had a reduced set. This only works in a limited situation. Gregor Hagedorn responded that this was only a simple example. Jessie Kennedy thinks you can't limit too much. Also, not names – concepts! There followed a detailed conversation with respect to what was needed from the interfaces, the mechanisms and possible limited usability, LSIDs and/or more user-readable keys. Gregor Hagedorn considers the minimum interface is the ID with no mandatory elements.

There was a long discussion concerning GUIDs, the different protocols and the combination of DC and ABCD.

Stan Blum – is DC an assembly of interfaces to complex objects? Gregor Hagedorn – yes, you are extending though composition. Jessie Kennedy also likes the idea of modularising but she does think we need an assembly of well-defined schemas not just interfaces. This should then result in a “schema” that encompasses all there is to say about a specimen, a concept, whatever. Flattening everything out is a detail of application.

Gregor Hagedorn – think of it terms of a semi-transparent abstraction layer beneath the top layer object model. Jessie Kennedy worries whether this is confusing.

Greg raised the point of having global registries (SDD, TCS, etc.)

There followed an ad hoc session for discussion of the Literature module

14 October 2004

Morning Session

TDWG Business – Chair: Stan Blum

Treasurer’s report

Walter Berendsohn apologised for a minor typographic error.

Walter Berendsohn explained events throughout the year, that not all accounts from Oeiras were in and that this year Secretarial travel costs have been paid because the NHM was unwilling to provide funding.

TDWG 2005

We have received an invitation from Alexander Ryss at the Zoological Institute, St Petersburg, Russia. Weather considerations suggest that the meeting should be held in September. Avoiding family holidays we suggest the week beginning 12 September 2005.

Reports from Subgroups

Economic Botany

A short report was received detailing progress towards a new standard.

Election of Officers

Jim gave a short presentation on the rules governing the election of officers and duration of posts.

His slate of nominations for Chair, Treasurer and Secretary was approved by a show of hands with no dissention.

Chairman Report

Stan Blum said he thought the 2004 meeting has been exceptional, much more than simple reports back from subgroups. He expressed his admiration for all involved.

Statistics: at the time of speaking there were 113 delegates registered and present from 86 organisations in 26 countries.

AOB

Walter Berendsohn: thanked Stan Blum for the past three year's strong leadership and the work done behind the scenes.

Contributed Papers (Taxonomic Names) – Chair: Jim Beach

Charles Hussey gave a presentation on Using informal names within taxonomic datasets; a practical perspective.

There were some questions on the procedures and human resource costs.

Yde de Jong gave a presentation on Authorship: Zoological versus botanical nomenclature.

Walter Berendsohn – the major difference in botany is the species name is an epithet whereas in zoology one can use the name. There was some disagreement about this in the supply of data to Fauna Europea. Yde said it was not possible to allow both methods in the one system and that was why the botany system was adopted.

Gregor Hagedorn – said it was essential to be able to resolve the ambiguities resulting from the different codes but this should be at the reporting, not data level.

Nico Franz gave a presentation on Towards a language for indicating relationships among taxonomic concepts.

There was short discussion especially on the mechanisms, the use of ranks, ambiguity.

Brook Milligan presented Taxonomic Names Meet Formal Parsers: Building Better Biodiversity Databases

Discussion: why can't you use XML? We have a lot of free text strings that would have to be marked up.

Marc Geoffroy presented THE TRANSMISSION ENGINE: accessing biological content linked to names representing different taxonomic concepts

Rich Pyle presented Protonyms, References, and Assertions: An introduction to the Taxonomer data model

James Ytow presented Formal Taxon Concept and Rough Set Approximation

Lunchtime

Computer Demonstrations and TDWG Executive Meeting

Afternoon Session – Chair: Stan Blum

Dave Thau – Ontogenies and OWL, with some notes on XML Schema, E-R Diagrams and UML in 15 mins

Anna Weitzman & Chris Lyal - taXMLit: a vital piece of the puzzle for digitally interoperable taxonomy

Anna Weitzman & Chris Lyal - Data standards: objective data, subjective data, and data interchange

Javier de la Torre - A world of conceptual schemas: mapping of standards to databases and other standards

Dag Terje Endresen - Genebanks as GBIF data providers – first experiences

Patricia Mergen - Mapping of OECD-MDS (culture collections standard) to ABCD

Mark Costello - Progress in, and plans for, the development of the Ocean Biogeographic Information System, including the need for standards to enable mapping marine species over the internet

Allen Rodrigo - www.DNA-surveillance: Web-based Genetic and Phylogenetic Identification of Taxa -- tools, methods and future developments

Patricia Koleff - Repatriation of Specimen Data to Reinforce National Systems of Biodiversity Information

15 October 2004

Morning Session

Contributed Papers – Chair: Reed Beaman Beeman

Eduard Stloukal - Towards indexing the biodiversity of the Carpathian region (Europe)

Giorgos Ksouris - Use and Functions of GBIF UDDI Registry

Robert Guralnick & Dave Neufeld - Research Challenges in Using Distributed GIS Services to Support Biodiversity Visualization and Analysis

Nelson Rios - A New Computing Tool for Automated Georeferencing of Natural History Collection Data

Paul Schreilechner - Implementing Open Geospatial Consortium standards for the GBIF-Austria web portal

Peter Brewer - Problems and some solutions to modelling species distributions at a global scale: the importance of taxonomic and specimen databases

Neil Caithness - BDWorld: A grid-based workflow manager for high-throughput distributed computing in biodiversity research

Tim Sutton - openModeller - An open, collaborative environmental niche modelling toolkit

Chuck Miller - Tropicos: *The Next Generation* - Lessons Learned on the Biodiversity Information Highway

Alan Paton - iPlants - Procedures for Collaborative Data Gathering

Hannu Saarenmaa [for Annie Simpson] - A call for TDWG support in the creation of invasive alien species standards

Doug Fils - *CHRONOS* : A federation of databases to support taxonomic data preservation for the paleobiological community

Phillip Dibner - Geospatial Interoperability Standards and Emerging TDWG Specifications

Lunchtime

Computer Demonstrations

Afternoon Session – Chair: Stan Blum

Karl-H Lampe - Digitized Insect Specimen Access: new digitization tools

Austin Mast - MorphBank: Web Image Database Technology for Biology

Arturo Ariño - Specimen Image Databases for Taxonomic Research: Experiences on File and Metadata Handling

Bruce Stein - Integrating Observational Data into Biodiversity Information Networks

Nancy Jacobsen Stout - Initial dissemination of MBARI's unique deep-sea biodiversity data

Robert Morris - Network access to XML data embedded in JPEG2000 images

Final TDWG Business

Stan Blum's summary:

Really pleased about work done to start to bring DiGIR and BioCAsE together.

Also impressed by some of the things Gregor Hagedorn introduced during the UBIF session.

Rebecca Shapley was asked to show the meeting her take on what had come together during the sessions, and a framework for a GBIF Schema Registry.

Stan Blum expanded on what Rebecca Shapley had showed us, referring to the different levels of complexity each schema and how they might relate and be combined. Maybe we shouldn't be thinking about extensions but adapting to different levels of complexity. Chuck Miller made the comment that we should sort out the obvious, low level conflicts (label conflicts, etc.). He (sort of) proposed a supervisory body. Javier said we should get as much as possible in the "same place". Bob Morris asked if we need to register ontologies and well as schemas. Donald Hobern said he can define what goes into the Registry.

Rich said when one domain bumps into another one you map them across with pointers, avoiding duplication. Jessie Kennedy agreed but warned against one group putting something in their schema that really belongs in someone else's. Anna Weitzman said it had been a great opportunity to see all that was going on.

Adrian said that we needed some sort of overarching mechanism within TDWG to manage the process, to avoid duplication, etc. Stan Blum said this is the function of his proposed Steering Committee. Rebecca Shapley gave a similar view. Charles Copp said that some of the items on Rebecca Shapley's chart were non-linear. There was much further discussion on this subject.

Jessie Kennedy said Darwin Core is not the representation of a specimen. [Unknown Speaker] said that there is not a very high enough standardisation across schemas (elements vs attributes). Stan Blum and Gregor Hagedorn said that they had looked at various ways of modelling. Schemas were the best way and guidelines existed.

Final Business & Wrap-up:

Stan Blum said he is amazed by the standard and number of presentations this year and praised all involved.

Minutes of last meeting – accuracy & corrections – Stan Blum advised of a typographic error and one error of fact.

Stan Blum listed the new state of TDWG Officers:

Chair: Walter Berendsohn
Treasurer: Stan Blum
Secretary: Adrian Rissoné

Regional Secretaries:

Africa: not represented
Asia: Yozomi (James) Ytow
Latin America: Patricia Koleff Osorio
North America: Gerald "Stinger" Guala
Oceania: Alex Chapman

Stan noted that Newsletter Editor remains vacant. Please twist arms!

New sub/interest groups:

Observations
Natural Collections Descriptions
Images
Bibliographic References & Taxonomic Literature

Stan Blum described changes in the way subgroups will be expected to work. A "charter" will be posted on the web site. There will also be a web site archive.

Next meetings (provisional): St Petersburg Sep 11th-19th 2005
2006 New Orleans, mid-Oct.

To warm applause, Stan Blum thanked the whole Landcare crew for their hard work and hospitality:

Close of Plenary

