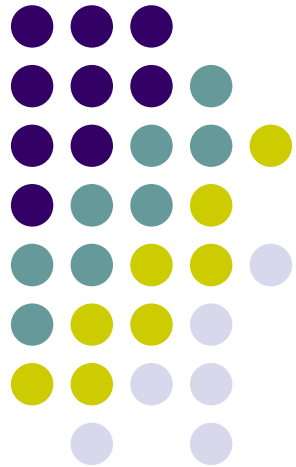
New Standards from Old:
reconciling HISPID with ABCD

Peter Neish
National Herbarium of Victoria; Royal Botanic Gardens Melbourne

Ben Richardson
Western Australian Herbarium; Department of Environment and Conservation; Perth

Greg Whitbread
Australian National Herbarium; Centre for Plant Biodiversity Research; Canberra

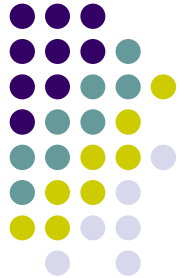
Biodiversity Information Standards (TDWG)
Bratislava, September 2007



Outline

- What is HISPID
- Why we needed to do something
- Approach to bringing standard up to date
- Outcomes

History of HISPID



- HISPID1 1989
- HISPID2 1993
- HISPID3 1996 (ratified by TDWG)
- HISPID4 2000 (web version only)
- HISPID5 in progress…

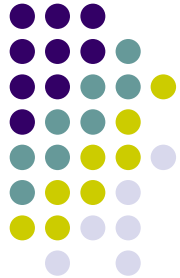
What is HISPID

Herbarium Information Standards and Protocols for Interchange of Data

- Flat file (ASN.1) format
- Non XML
- About 150 concepts
- Controlled Vocabularies
- Primary use case: transferring data with specimen exchange
- Developed by the Australian Herbarium Information Systems Committee (HISCOM)

```
startfile
version         "HISPID3",
numrec          2,
datefile        19951202,
institute       "National Herbarium of New South Wales (NSW)",
contact         "Gary Chapple",
address         "Royal Botanic Gardens, Mrs Macquaries Road, Sydney NSW 2000, Australia",
phone           612 92318164,
fax             612 92517231,
email           "gary@rbgsyd.gov.au",
fileaction      "insert",
filedescriptor  "exchange",
content         "Herbarium exchange data of various species from NSW to CANB",
{
insid           "NSW",
accid           "390839",
fam             "Loranthaceae",
gen             "Amyema",
sp              "pendulum",
isprk           "subsp.",
isp             "longifolium",
vnam            "Wiecek, B.M.",
vdat            1995,
prot            "Wild",
cou             "AUSTRALIA",
pru             "NSW",
sru             "Central W. Slopes",
loc             "Mount Bolton, Moura",
```

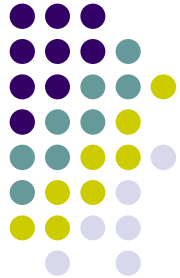
Specimen duplicates (replicates) use case



Collection of more than one sample from a single plant from the same locality at the same time.

- Multiple physical and digital objects, but tightly coupled (almost all data the same)
- Up to one third of specimens will have at least one replicate in another herbarium
- Significant opportunity to reduce data entry costs
- To make full use of duplicate data, you need to be about transfer **full** database records

What's wrong with HISPID?



The Good

- It works! (for data transfer)
- Well defined concepts with community agreement
- Agreed controlled vocabularies
- Simple to implement

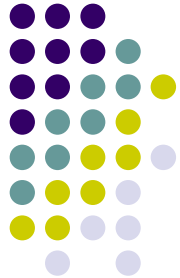
The Bad

- Flat file structure
- Lack of mechanism to enforce syntax - relied on convention and documentation

The Ugly

- Not suitable to web applications and new use cases

Australia's Virtual Herbarium (AVH)



- Initially used HISPID3 (internally converting to XML)
- Then moved to BioCASe protocol and ABCD
- Now implementing TAPIR

But couldn't transfer full HISPID records:

- ABCD was missing concepts from HISPID
- ABCD controlled vocabularies differed from HISPID
- Multiple places to put information
- Complexity = more difficult to implement

Some examples

HISPID Concepts not in ABCD

- Frequency
- Introduction Agency
- Life Form
- Number of Sheets
- Phenology
- State
- Secondary Collector Identifier
- Soil Type
- Substrate

…

HISPID Controlled Vocabularies not in ABCD

- Altitude Measured By
- Coordinate Method
- Cultivated Occurrence
- Depth Measured By
- Type Qualifier
- Identifier Role
- Kind of Unit
- Name Addendum
- Natural Occurrence
- Type Status

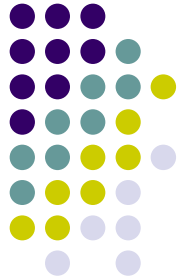
…

Reconciling HISPID and ABCD

Aims

- Use HISPID as a basis for adapting ABCD in a way that met requirements of the Australian Herbaria Federation
  - Use built in ABCD mechanisms to extend schema
  - New schema would be 100% valid ABCD
  - Keep changes manageable and maintainable into the future

First Attempt

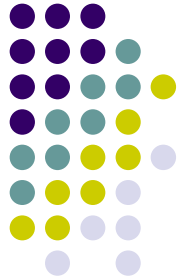


- Used the generalised ABCD concepts:
  - Gathering/SiteMeasurementsOrFacts
  - Gathering/Biotope/MeasurementsOrFacts
  - Unit/MeasurementsOrFacts

But:

- didn’t allow for validation
- reliance on conventions – lose the benefits of XML Schema validation
- consumers of data would have to know in advance the MeasurementOrFact/Parameter

Second Attempt



- Using the extension mechanism of ABCD
  (like ABCDEFG and ABCDEFGHHI)
  - Allowed us to create our own elements with controlled vocabularies and repeating elements
  - Allowed us to redefine existing ABCD types to better fit HISPID concepts

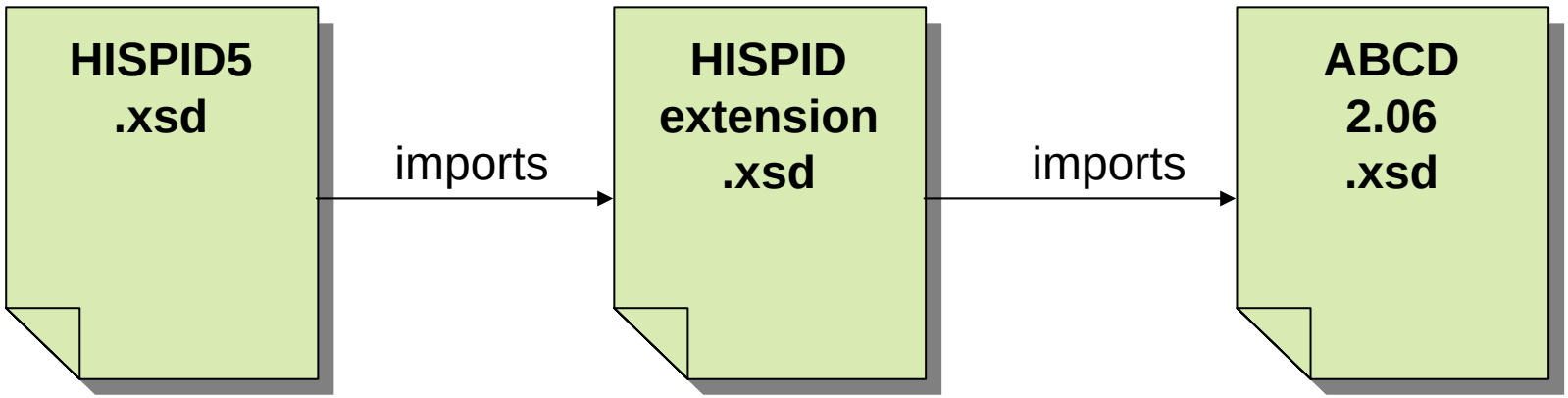
But:

- Data in the extension is effectively hidden from consumers outside the federation
- Redefining an existing ABCD type was overkill if all we wanted to do was restrict the allowed values

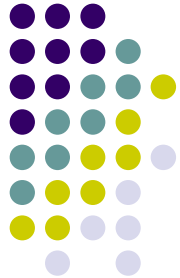
Third Attempt

Two parts:

- Concepts not in ABCD placed in Extension
- Used XML:Schema to restrict existing ABCD concepts to HISPID vocabularies



Some example new concepts



Uncontrolled:

- IntroductionAgency
- LifeForm
- NameFormula
- NumberOfSheets
- PrimaryRecordingUnit
- Substrate

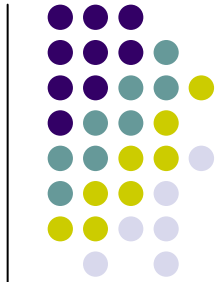
…

Controlled:

Phenology

- bisexual flowers
- buds
- female cones
- female flowers
- flowers
- male/female cones
- male cones
- male flowers
- fruit
- fruiting cones
- gametophyte
- sporophyte
- spore-bearing bodies
- fertile
- sterile
- leafless

Example restriction



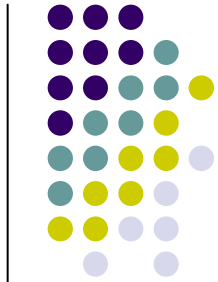
CoordinateMethod

ABCD type: abcd:String

HISPID type: hispid:CoordinateMethodEnum

- Restricted with xs:enumeration to:
  - `collector`
  - `compiler`
  - `generalised arbitrary point`
  - `automatically generated`
  - `gps`
  - `topo`
  - `unknown`

Problems still to resolve



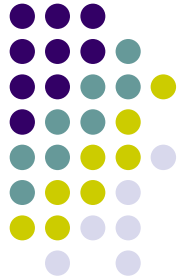
Many examples where extending existing ABCD elements would have made more sense than adding new elements e.g.:

- Adding some additional values to existing ABCD enumerations (e.g. Ordination, HigherTaxon)
- Adding attributes to exisitng ABCD elements (e.g. “per” collector attribute to GatheringAgent)
- Adding additional elements to existing ABCD types (e.g. DegreesLatLong type – add in elements for DMS)

But:

- Adding new things to this schema outside of the extension placeholder would have created invalid ABCD, so either placed in extension or not implemented but documented for consideration

Problems still to resolve

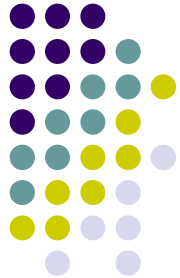


- Anonymous types in ABCD are harder to extend, therefore harder to maintain

  (goal was to keep HISPID in synch with ABCD)

  - E.g. IdentificationQualifier – would like to restrict this to a controlled vocabulary:
    - `aff.`
    - `cf.`
    - `incorrect`
    - `forsan`
    - `near`
    - `?`

Problems still to resolve



- Single ABCD types that are used in different contexts throughout schema are hard to restrict without creating a new type for each context

  e.g. restrictions for

  Altitude/MeasurementOrFactAtomised/Method
  - `dem`
  - `gps`
  - `field estimate`
  - `altimeter`
  - `map`
  - `unknown`

  c.f. Depth/MeasurementOrFactAtomised/Method
  which would have different values

Outcomes

- A schema that works for the Australian Herbarium community
- Approximately 150 HISPID5 concepts documented as a useful subset of ABCD for herbaria
  - Most concepts map 1:1 HISPID to ABCD
  - 11 restrict ABCD concepts to HISPID vocabulary
  - 20 new concepts (but some could probably be incorporated into ABCD)
- Potential changes to ABCD documented – may or may not be implemented
- 100% ABCD compatible, so we can:
  - participate in global systems (eg GBIF)
  - simplify mapping HISPID to TDWG ontology

Acknowledgements

- Council of Heads of Australasian Herbaria
- Herbarium Information Systems Committee (HISCOM)
- TDWG Infrastructure Project