FACET ANALYSIS AS A TOOL FOR MODELLING SUBJECT DOMAINS AND TERMINOLOGIES

REPRESENTING FACETS ON THE WEB

Some questions and observations:

- there is a parallel between ‘topic’ as a property of a knowledge object and the way ‘subject’ is dealt with in a bibliographic record
- dealing with a KOS or a terminology is different from managing resources themselves
- do we need to deal with the properties of concepts in a different way from the properties of ‘knowledge objects’

General applicability of facet theory:

- facet analysis has some merit as a general methodology for modelling domains
- it has a proven track record for the creation of structured vocabularies
- it identifies a wide range of attributes and relationships between concepts (and has the capacity to do more)
- the logical nature of the analysis (and the structure of resulting systems) makes it compatible with automation and susceptible to machine manipulation

Facet analysis as a generalised modelling tool:

- originally envisaged as a means of reducing complex subject content to a predictable linear order for physical organization
- facet analysis achieves four broad objectives:
  1. it categorizes concepts into functional groups
  2. it imposes order between concepts
  3. it identifies relationships between concepts
  4. it provides a system syntax for managing combination in the case of complexity

  to some degree it shows the features of an ontology

Conventional tools based on facet analysis:

- application of the general methodology produces logical and well ordered structures
- internal organization of facets is straightforward
- synthesis of concepts within and between facets is easily and predictably managed
- highly sophisticated levels of organization can be achieved without compromising the underlying principles
- examples in this presentation use the recently revised Class C Chemistry of the 2nd edition of the Bliss Bibliographic Classification (BC2)

. . *Kinds of catalysis & catalysts*

. . . *By physical location*

N . . . . Fixed catalysts

O . . . . Mobile catalysts

. . . *By phase conditions*

P . . . . Homogeneous catalysis

* Catalyst and reactor are in the same phase.

PS . . . . . Acid-base catalysts

Q . . . . Heterogeneous catalysts

* Catalyst and reactor are in different phases.

. . . *By effect on composition*

R . . . . Physical catalysts

* Add to CCA R letters A/W following B; eg

RBJ . . . . . Pressure catalysts

S . . . . Chemical catalysts

* Usually assumed.
* For reaction product as catalyst, see autocatalysis CCA U.
* See also Enzymes CUL

TB . . . . . Mixed catalysts

. . . . . Particular substances

TH . . . . . Water (as catalyst), hydrolysis

Figure 1: *Internal organization of a facet*

CNO MJT	Compounds with oxygen & other metals
MJT J	. Mixed metal oxides of iron
	* For specific metals, see the metal; eg, Strontium ferrate (Sr2FeO4) CLA NOM IFN.
MK	. Compounds with oxygen & hydrogen
MKJ	. . Hydroxides
MKJ HL	. . . Ferrous hydroxide, iron (II) hydroxide
MKJ HN	. . . Iron (III) hydroxide
MQ	. Compounds with sulphur
MQJ	. . Sulphides
MQJ HL	. . . Ferrous sulphide
MQJ HP	. . . Ferrous disulphide
MQM	. Compounds with sulphur & oxygen
MQM IFL	. . Ferrous sulphate, iron (II) sulphate
	. . . Hydrate
MQM IFL KMP	. . . . Heptahydrate ferrous sulphate, green vitriol, copperas
MQM IFN	. . Ferric sulphate, iron (III) sulphate
MV	. Compounds with chlorine
MVJ	. . Chlorides
MVJ HL	. . . Ferrous chloride, iron (II) chloride
MVJ HN	. . . Ferric chloride, iron (III) chloride

Figure 5: *Expansion of schedule through combination of concepts within facets*

Kinds of chromatography

A . In combination with other analyses

. *By mobility of phase*

B . . Stationary phase (chromatography), fixed phase (chromatography)

BL . . . Solid support (chromatography)

* When stationary phase is liquid.

C . . Moving phase, mobile phase

CG . . . Carrier gas (chromatography)

. *By process involved, separation mechanism*

* For electrochromatography, see C9Q H.

D . . Adsorption chromatography

D7O G . . . Activation

Figure 6: *Combination of concepts across facets*

. . *Operations & agents*

YQ . . . Organization & management of scientific work

* As TQ (including operations research).

AY3 2 . . . Research operations (general)

2C . . . . Methodology...

4 . . . Theory

* Alternative (not recommended) to locating after Practical science, at AY8 B.

6 . . . Practical scientific work

7 . . . . Unwanted effects... Safety & protection...

. . . . *Agents*

B . . . . . Equipment & materials (together)

. . . . . *Operations on equipment & materials*

C . . . . . Handling techniques...

F . . . . . *Processes in equipment & materials*

G . . . . . Deterioration...

J . . . . . *Properties of equipment & materials*

JE . . . . . Reliability... Responsiveness...

K . . . . . *By energy interactions & forms*

* Add to AY3 K letters A/V following B Physics; eg magnetic properties AY3 KJ.

. . . . . *Parts*

NB . . . . . Surfaces

. . . . . *Types of equipment & materials*

PD . . . . . Smart equipment & materials...

Figure 3: *Repeating citation patterns in BC2*

Broader applications of faceted terminologies:

- these attributes of a faceted terminology produce structures that are not only complex and semantically rich, but also logical
- they also support output of the terminology in various formats
  1. conventional classification
  2. thesaurus format
  3. (potentially) ontological structures
- previous work on the second edition of the Bliss Bibliographic Classification (BC2) has concentrated on the machine generation of 1 and 2
- current work is investigating how the faceted structure can be represented in a web format

BC2 Class C – Chemistry:

- chemistry is theoretically logically structured, but there are practical difficulties in the representation of its considerable complexity
- examples will show how the basic vocabulary is encoded for output as a classification or thesaurus and demonstrate the way in which the BC2 software operates

BC2 markup language:

- BC2 already exists in a machine readable form
- it uses a simple encoding system that enables machine inference of a number of structural features of the classification
- the encoding identifies:
  1. hierarchical position
  2. non-classes such as principles of division or other ‘signposts’
  3. status of a class for inclusion in the index
  4. formatting of index entries

```
CCA             05Catalysis, catalysts
@               06(Operations on catalysts)
CCAG            07Catalyst carrier
CCAH            07Catalyst stripping
CCAI            07Poisoning of catalysts, anti-catalysis
CCAJ            07Regeneration of catalysts
@               06(Parts of catalysts)
CCAL            07Action centre (catalysts)
@               06(Kinds of catalysis & catalysts)
@               07(By physical location)
CCAN            08Fixed catalysts
CCAO            08Mobile catalysts
@               07(By phase conditions)
CCAP            08Homogeneous catalysis
                     *   Catalyst and reactor are in the same phase.
CCAPS           09Acid-base catalysts
CCAQ            08Heterogeneous catalysts
                     *   Catalyst and reactor are in different phases.
@               07(By effect on composition)|
CCAR            08Physical catalysts
                     *   Add to CCA_R letters A/W following B; eg
CCARBJ          09Pressure catalysts
CCAS            08Chemical catalysts
                     *   Usually assumed.
                     *   For reaction product as catalyst, see
                         autocatalysis CCA_U.
                     *   See also Enzymes CUL
CCATB           09Mixed catalysts
@               09Particular substances ]IT
CCATH           10Water (as catalyst), hydrolysis
CCATJ           10Others ]IT
                     *   Add to CCA_T letters J/S following C for
                         general substances.
                     *   Add to CCA_T letters T/Y for catalysts in a
                         special context; eg Peptidases CUF_CAT_T.
```

Figure 8: *Coded classification data in source file for Class C*

Generation of schedules and index:

- from this basic input data the software can infer the hierarchical display and other layout elements for publication as a classification schedule
- it will also generate an alphabetical index, although this still requires some manual editing
- note that the index is derived from the classification itself and not built independently
- we know that the source code could be further developed to increase automatic reasoning
- for example if facet/category status were encoded some classes of associative relationships could be inferred, and automatic number building supported

CCA . Catalysis, catalysts

. . *Operations on catalysts*

G . . . Catalyst carrier

H . . . Catalyst stripping

I . . . Poisoning of catalysts, anti-catalysis

J . . . Regeneration of catalysts

. . *Parts of catalysts*

L . . . Action centre (catalysts)

. . *Kinds of catalysts & catalysts*

. . . *By physical location*

N . . . . Fixed catalysts

O . . . . Mobile catalysts

. . . *By phase conditions*

P . . . . Homogeneous catalysis

- Catalyst and reactor are in the same phase.

PS . . . . . Acid-base catalysts

Q . . . . Heterogeneous catalysts

- Catalyst and reactor are in different phases.

. . . *By effect on composition*

R . . . . Physical catalysts

- Add to CCA R letters A/W following B; eg

RBJ . . . . . Pressure catalysts

S . . . . Chemical catalysts

- Usually assumed.
- For reaction product as catalyst, see autocatalysis CCA U.
- See also Enzymes CUL

TB . . . . . Mixed catalysts

. . . . . Particular substances

TH . . . . . Water (as catalyst), hydrolysis

TJ . . . . . Others

- Add to CCA T letters J/S following C for general substances.
- Add to CCA T letters T/Y for catalysts in a special context; eg Peptidases CUF CAT T.

Thesaurus generation:

- the same input data will also generate a thesaurus
- this involves the same methodology as the manual derivation of a thesaurus from a faceted classification
- the software replicates the intellectual process of identifying relationships between concepts through the examination of the classification structure
- it is able to infer both equivalence relationships and narrower term/hierarchical relationships and their reciprocals
- knowledge of the scheduling rules of the classification is built into the program

BJ Magnetism

* See note at BGY re magnetism and bulk matter.
* Add to BJ letters A/Y following BVJ where applicable; eg

BJB H . Magnetic field

. *Processes & properties*

IM . . Magnetic moment, dipole moment

BJC P . . Relaxation

BJF O . . Resonance

BJK . . Magnetic flux

S . . . Susceptibility

BJL . . Magnetization

. *Field components*

BJN . . Magnetic monopoles

Q . . Dipoles

. *Forms of magnetism*

BJQ . . Diamagnetism

BJR . . Paramagnetism

BJU . *Interactions with other energy forms*

B . . Magnetomechanical effects

C . . . Magnetostriction

Figure 7: *Thesaural relationships exposed in the faceted display*

Structural and syntactic rules:

- the software has the following information about any individual class derived from the markup

1. its position in the hierarchy of items
2. its classmark, if it has one
3. its names
4. its cross-references, if it has any
5. its importance indicator (a device to ensure appropriate column/page breaks), if it has one.

Structural and syntactic rules:

- the program also has knowledge of sequencing in the classification based on rules of syntax:

1. a class
2. its first offspring
3. if no offspring, its next sibling
4. if no more siblings, its parent's next sibling
5. if no more parent's siblings, its grandparent's next sibling
6. and so on, ad infinitum.

Program output:

- from this data the software can infer the relationship between pairs of classes
- it labels them accordingly to produce the thesaurus format
- although some manual editing is required, no intellectual input is necessary for this process to occur
- if the structure of classification is correctly established and encoded, the thesaurus will sit on the back of it
- this phenomenon confirms the applicability of facet analysis beyond the limits of classificatory structures

Neo-impressionism (1880s/1890s) WK9 FLT N
BT 9th decade WK9 FLT
NT Pont Aven School WK9 FLT NP
Neolithic WC7 RH7 8X
BT Aegean art WC7 RH
NT Sesklo period WC7 RH7 8XS
Neolithic WDR 78X
BT Chinese visual arts WDR
NT Liangzhu culture WDR 78Y LI
NT Longshan culture WDR 78Y LO
Neolithic WDS 78X
BT Prehistory WDS 78
Neolithic art WC7 8X
UF New stone age art
BT Stone age art WC7 8B
NT Animal style (Neolithic art) WC7 8X3 SAN
Neolithic art WC7 P78 X
UF New stone age art
BT Stone age art WC7 P78 B

Netherlands art WCJ U
UF Dutch art
BT European visual arts WCD
NT 20th century WCJ UN
Netherlands painting WK9 JU
USE Dutch painting
Netsuke WHG VN
BT Articles of adornment WHG T
Netting WGT SU
BT Lace work WGT S
NT Egyptian network WGT SUW
NT Filet WGT SUX
NT Tulle WGT SUU
Networks WTU CP
USE Controllers
Neue Sachlichkeit WCK NGN
UF Magic realism
UF New objectivity
BT 1920s WCK NG

Figure 10: *Thesaurus display generated from source code in BC2*

Vocabulary control in BC2:

- lack of editorial policy in the drafting of early schedules throws up some difficulties in the area of vocabulary control
- many of the class names are not suitable as thesaurus terms
- formatting of class names is not consistent
- preferred terms are not clearly indicated
- this means that existing schedules require heavy editing for the thesaurus format
- better editorial control of future schedules will make the process more easily managed

```
JGD             04^Student organisations, societies ]T, clubs ]T, student
                        societies= student clubs=clubs, student=student
                        societies= societies, student ]SI
                    *   Organisations based on institution attended.
JGE             05Officers ]T= Officers of student organisations
JGEH            06Prefects, monitors, form leaders ]T
JGEK            06Heads of schools ]T
JGEP            06Presidents, vice-presidents ]T
JGES            06Secretaries ]T
§               05(Particular organisations)
                    *   Those serving a particular activity are
                        subordinated to the activity.
JGF             06Students' unions
JGG             06Fraternities=, sororities=, Greek letter societies
JGJ             06Alumni=, alumnae=, graduate organisations
JGK             05Social activities ]T=, cultural activities ]T= ^Students
social
                        Activities= students cultural activities ]SI
JGL             06Journalism ]T=, student magazines ]T= Student journalism=
                    journalism, student= student magazines ]SI
```

Figure 11: *Editorial revision for thesaurus generation*

Representing facets on the web:

- in order to interact fully with the semantic web, a faceted terminology must be visible there
- more importantly, all the aspects and functions of a faceted system must be visible too
- the current challenge for BC2 is to see how this can best be achieved
- BC2 is at present only visible as:
  - PDF files of drafts as Word documents
  - PDF files of camera ready copy of published classes derived from source code
  - some limited examples of source code

	Introduction and auxiliary schedules	1977 * ISBN: 0408708212
2/9	Generalia, phenomena, knowledge, information science & technology	
A/AL	Philosophy and logic	1991 ISBN: 1857390253
AM/AX	Mathematics, probability, statistics	1993 ISBN: 1857390725
AY/B	General science, Physics	1999 ISBN: 0408708247
C	Chemistry (Unpublished incomplete draft)	
D	Astronomy and earth sciences	
DG/DY	Earth sciences (Unpublished draft available in pdf format)	
E/GQ	Biological sciences (Unpublished incomplete draft)	
GR/GZ	Applied biological sciences: agriculture and ecology	
H	Physical anthropology, human biology, health sciences	1980 * ISBN: 040870828X
I	Psychology and psychiatry	1978 * ISBN: 0408708417
J	Education (Rev. ed.)	1990 ISBN: 0862912784
K	Society (includes social sciences, sociology and social anthropology)	1984 * ISBN: 0408708301
L/O	History (including area studies, travel and topography, and biography) (Two unpublished drafts available in pdf format, by Haddon Library and Sidney Sussex College)	
LA	Archaeology (Unpublished draft available in pdf format)	
P	Religion, occult, morals and ethics	1977 ISBN: 0408708328
Q	Social welfare and criminology (Rev. ed.)	1994 ISBN: 1857391217
R	Politics and public administration	1996 ISBN: 1857390776
S	Law	1996 ISBN: 1857390679
T	Economics and management of economic enterprises	1987 ISBN: 0408708344
U/V	Technology and useful arts (including household management and services)	
W	The Arts (Available on this site in pdf format)	2007 ISBN: 3598243359
WV/WX	Music (Unpublished draft available in pdf format)	
X/Y	Language and literature (Unpublished draft available in pdf format)	
ZA/ZW	Museology (Unpublished draft available in pdf format)	

Options for future development:

- the current coding system requires the use of the customised software to make it work
- it could be converted to another format
- alternatively future schedules could be encoded entirely differently
- in that case the optimum format must be decided
- should we regard BC2 as:
  - a text
  - a database
  - an ontology

Existing formats:

- a form of XML exists for faceted tools (XFML)
- but it is no longer supported
- it is relatively simple and doesn't look particularly compatible with BC2

City (City is a facet)

- New York (New York is a topic within the facet City)
- L.A.

Type of place

- Bars
- Restaurants

Type of music

- Blues
- Latin

```
<facet id="city">City</facet>
<facet id="place">Type of place</facet>
<facet id="music">Type of music</facet>
<topic id="ny" facetid="city"><name>New York</name></topic
<topic id="la" facetid="city"><name>Los Angeles</name></tc
<topic id="bar" facetid="place"><name>bar</name></topic>
<topic id="restaurant" facetid="place"><name>restaurant</r
<topic id="blues" facetid="music"><name>blues</name></topi
<topic id="latin" facetid="music"><name>latin</name></topi
```

Figure 12: *Representation of faceted structure in XFML*

Converting BC2 source code to XML:

- this proved surprisingly easy to manage
- a simple program achieved output of the code as XML
- it identifies structural features of the BC2 KOS such as principles of division and scope notes
- but it looks very much like a digital text
- not entirely clear whether it preserves the hierarchical structure
- it almost certainly lacks the functionality of the original code + software in terms of the automatic reasoning

```
- <concept id="CCA" term="Catalysis, catalysts">
  - <charOfDivision term="(Operations on catalysts) ]IT">
      <concept id="CCAG" term="Catalyst carrier" />
      <concept id="CCAH" term="Catalyst stripping" />
      <concept id="CCAI" term="Poisoning of catalysts, anti-catalysis" />
      <concept id="CCAJ" term="Regeneration of catalysts" />
    </charOfDivision>
  - <charOfDivision term="(Parts of catalysts) ]IT">
      <concept id="CCAL" term="Action centre (catalysts)" />
    </charOfDivision>
  - <charOfDivision term="(Kinds of catalysis & catalysts) ]IT">
    - <charOfDivision term="(By physical location) ]IT">
        <concept id="CCAN" term="Fixed catalysts" />
        <concept id="CCAO" term="Mobile catalysts" />
      </charOfDivision>
    - <charOfDivision term="(By phase conditions) ]IT">
      - <concept id="CCAP" term="Homogeneous catalysis">
          <scopeNote>Catalyst and reactor are in the same phase.</scopeNote>
          <concept id="CCAPS" term="Acid-base catalysts" />
        </concept>
      - <concept id="CCAQ" term="Heterogeneous catalysts">
          <scopeNote>Catalyst and reactor are in different phases.</scopeNote>
        </concept>
      </charOfDivision>
    - <charOfDivision term="(By effect on composition) ]IT">
      - <concept id="CCAR" term="Physical catalysts">
          <scopeNote>Add to CCA_R letters A/W following B; eg</scopeNote>
          <concept id="CCARBJ" term="Pressure catalysts" />
        </concept>
      - <concept id="CCAS" term="Chemical catalysts">
          <scopeNote>Usually assumed.</scopeNote>
          <RT idref="CCAU">For reaction product as catalyst: Autocatalysis</RT>
          <RT idref="CUL">Enzymes</RT>
          <concept id="CCATB" term="Mixed catalysts" />
        - <charOfDivision term="Particular substances ]IT">
            <concept id="CCATH" term="Water (as catalyst)" />
```

Figure 13: *XML output from BC2 source code*

A skos form of BC2?

- If we look at the skos elements, they bear a better relationship to BC2 encoding
- skos can represent:
  - editorial elements
  - structural elements
  - some relationships
  - hierarchical relationships
  - equivalence relationships
- in some respects it is more specific than BC2
- but some aspects of BC2 are missing

SKOS elements	BC2 elements	Encoded as
skos:concept	BC2 class	(0x)
skos:broader	BC2 superordinate class, BT	(0x+1)
skos:narrower	BC2 subordinate class, NT	(0x-1)
skos:prefLabel	BC2 preferred term (lead term in class, index entry term)	
skos:altLabel	BC2 non-preferred term, synonyms and near-synonyms, non-lead terms	
skos:notation	BC2 classmark	e.g. CCA ...
skos:note	BC2 general notes, explanations	@
skos:editorialNote	BC2 instruction	@
skos:scopeNote	BC2 scopenote	@
skos:related	BC2 related term	RT
skos:broaderTransitive	BC2 super-superordinate classes	(0x+2)
skos:narrowerTransitive	BC2 sub-sub-ordinate classes	(0x-2)
skos:collection	BC2 facet or array, facet name or characteristic of division	@ ()
skos:orderedCollection	BC2 ordered facet or array (default position in BC2)	
skos:member	BC2 class in array	

Figure 13: Correspondence between SKOS elements and BC2 code

BC2 as skos:

- a skos version of BC2 would not be the same as the existing terminology
- it would not be a ‘thing’ in itself usable as a terminology
- it seems unlikely that we could export the BC2 code to a skos format as we did with XML
- it seems to imply a huge inputting effort (for which we don’t have the resource)
- would effort be better spent in setting BC2 up as a relational database (in the same way as UDC)

BCA www.blissclassification.org.uk