MINUTES FROM CONFERENCE CALL 24 Nov 05

Participating:

• Catherine Jones

• Cliff Morgan (Chair)

• Claire Saxby

• Andrew Wray

Claire, John, Andrew, Evan, Scott and Catherine had circulated use cases broken down by attribute values (in spreadsheet form).

Cliff asked how the group had been approaching this task – e.g. were we thinking of a specific fictional instance, and assigning values (such as IPR status, format, etc.) based on this fictional instance, or were we listing all possible values for the attributes (i.e. not thinking of a specific case), or just highlighting that we didn't know what the value was because we didn't have sufficient information to hand?. We agreed that the danger associated with inventing specific values is that this is fairly arbitrary – we will be influenced by our own preconceptions and will be ignoring other possible values. The resulting version descriptions would not be comprehensive and may be biased.

We agreed that some attributes, although interesting to discuss, may be limited in value. Also, in order to increase the likelihood that attribute values are included in a version's metadata, the number of attributes required should be as short as necessary. Andrew said that at this stage however we should retain all of the attributes in the spreadsheet, as we explore this further and identify whether different versions are adequately differentiated by their attributes.

Cliff led a discussion about which attributes might be most useful based on the exercise to date. He felt these could be grouped into three categories – i) attributes that could have a number of essentially free-form values, ii) those that can only have a fixed set of specific values, and iii) those that the exercise may reveal to have limited use, or are non-intuitive, or potentially confusing in their terminology.

i) ATTRIBUTES WITH "FREE-FORM" VALUES

These would include the following:

IP Ownership – complex and may vary over time as rights are transferred. Apart from author and publisher, IPR may be held jointly, or by the government, or by the employer of the author, for example. Or the rights may be specifically expressed in a Creative Commons licence. Because of this complexity, we felt that it would be better to allow for any copyright/Creative Commons type statement. (CM note following meeting: "IPR" also covers more than copyright – it could also cover patents rights, database rights, moral rights, typography rights.)

Identifiers – could be a specific document identifier or (as suggested in the attributes table) a citation.

Relationship – a key piece of metadata. This may change over time. Suggested that we look at how Dublin Core have identified in DC Terms.

Source – could have more values than listed in the table. Also, doesn't seem to be the right term: it's used to say where something is rather than where it comes from, so maybe "location" is a better term. (CM: "Source" is one of those terms that has benighted Dublin Core discussions over the years!)

Format –some material will be in more than one format, or have material embedded in it in one or more formats. Cliff cautioned against us going too far down the road of specifying format values and their various flavours. The set of metadata required to ensure a full description of material (e.g. for preservation purposes) is likely to be complex, and has already been the subject of other standards initiatives such as OAIS and Open E-Book, for example, who have identified the need to be very precise about every digital file, file-naming conventions, navigation through the material, digital manifests, etc. For our JAV group, we may just want to specify format at the highest level, and maybe reference MIME as a controlled list of format names?

ii) ATTRIBUTES WITH A LIMITED AND DEFINED SET OF POSSIBLE VALUES

Relationship may belong to this group instead of (i), depending on how comfortable we are about a limited set of relationship descriptions.

Fixity – Andrew suggested that 'Version of Record' replace the term "Final". If further changes are made to the Version of Record, they may be "Version of Record Updates", i.e. where the Update is associated with the Version of Record in some formal way, or they could be (for want of a better term) unauthorized Updates, where the material is amended but has not gone through any formal process (e.g. author updates version in IR; IR manager migrates article to a more recent format).

iii) ATTRIBUTES WITH SOME QUESTIONS RE UTILITY OR CLARITY

Bibliographic Context – not sure whether could be easier to represent under "Relationship"

Visibility – some confusion about what the 'General' value means. "Limited" means that access is restricted by the author or publisher, so this might be the equivalent of marking something "Private & Confidential". "General" seems to mean that the material is available for anyone, although the availability may be subject to certain conditions, e.g. anyone can subscribe as long as they pay the money. This is different from something being in the public domain. There may also be other specific restrictions to access, e.g. territorial (only available to people in particular countries) or age-related (only available to people over a certain age), etc. So we may need a fuller set of values, or maybe even put this into group (i).

Peer Review – Andrew pointed out that this is not a binary attribute, where material is either per reviewed or it isn't. There is the possibility of alternative quality

assessment models being used in the future that may be just as valid (or more so) than peer review. Also, the term is not such a great indicator of value since the quality of peer review itself varies enormously. Maybe delete from our attribute list?

Status – not clear how useful this was as a descriptor. Delete?

Delivery Context – ditto

Other comments from the meeting:

Claire had commented in her note with use case 13 (multiple authors submitting versions to different post-print servers which then diverge over time) that the way that the versions diverged over time, with possibly iterative author-amended and IR-manager-amended versions, was not described in the use case. Cliff said that in this case we would need to identify that there will be differences in the versions, and agree which attributes are most relevant to capturing these differences in the spreadsheet.

Cliff agreed with an e-mail comment from Evan that it would be sensible to include 'Not applicable' as a possible value for each attribute, since this is better than a blank field.

The ICMJE guidelines as raised by Scott were discussed. Claire said that these are frequently consulted by biomedical journals for e.g. good publication practice/ethics guidelines. We should check to see what the guidelines say about versions.

The next step is for everyone to complete their use cases, and to seek feedback from those unable to join the call regarding which attributes were found to be useful based on this exercise.

Minutes – Claire Saxby and Cliff Morgan