MINUTES OF MEETING OF JAV TECHNICAL WG 22 December 2005

Participating: Peter McCracken, Evan Owens, Clare Saxby, Bernie Rous

We had an informal conversation about the problem at hand and how we are going to make progress. We asked what is the next step and made a few attempts to answer that question.

Is it to trim the attributes to the most relevant and find common terminology to express that? If so, how are we going to do that? Is one person going to spend a lot of time wading through all the spreadsheets looking for patterns and distilling the information? Or, as Bernie suggested, should each person draw their own conclusions and then share those conclusions with the group.

We discussed some characteristics of the data that we have collected:

1) that all attributes are not yet fully normalized; some are really properties of related objects (a classic normalization technical problem) while others have "business" normalization problems, such as where two or more attributes have cross-dependencies for business reasons (e.g., "status" which is linked to IP ownership and Bibliographic Identity).

2) Bernie reported problems coping with IP ownership, particularly because it can change through events that are external to object at hand. This points at a larger problem that several of the attributes are not strongly fixed, but can change over time. Which in turn suggests that our data model is still too flat as pointed out in 1) above.

3) We discussed the sentiment expressed on earlier calls that relationship was the most important attribute. We considered whether relationship is an attribute of a version at all or better thought of as external property of a set of versions, in effect a query result rather than an attribute. Clare described her view of relationship, which was all the facts known at the time the version is created, in effect at a given point in time. We struggled a bit with whether relationship assumed a fixed point or definitive version or whether it was position within the entire universe of versions.

Bernie characterized our task as "to describe the ideal set of information that could be available with a version."

Bernie reported that the original DOI concept was that authors MSS should be assigned DOIs that follow through the life of the paper. He reported that idea got lost in the implementation of CrossRef but is now coming back a bit because CrossRef is talking with Institutional Repositories about

assigning DOIs.

We collectively agreed to urge the chair to propose next steps as we dive back into this problem in the new year.

****

NOTE FROM CLIFF MORGAN FOLLOWING MINUTES OF MEETING OF 22 December 2005:

Thanks very much Evan - and I note the urge for me to propose next steps.

My general feeling could be summed up as "keep it simple". The most specific help that we can give to the community is to propose a set of terms that we can use to describe JAVs. Whenever I speak with interested parties (most recently the Chair of the RCUK), the message that most strongly comes across is "please just say what we should call these various versions". Of course, the terms will need to have some definitions around them, but I don't think we should get over-involved in complex data models. (This reminds of the early Dublin Core discussions between minimalists and structuralists.)

I agree with Bernie's suggestion that each member can be asked to come up with his or her own conclusions from having done the spreadsheets - and I expect that we will have a spread of those who think (like me) that we should do some winnowing and those who think we should build a more robust data model.

Personally, I think the exercise has been useful in helping us to focus on the most meaningful/most knowable qualifiers that may be associated with a JAV. For me, these would be:

1. a shorthand description of the version (draft, etc.) - the term can include some reference to peer review if not self-defined within the term
2. an identifier
3. some statement about ownership and availability (i.e. encompassing both the Identifier and the Visibility qualifiers)
4. an open field for Relationship, in which links are given to all relevant other versions

I don't know how convinced I am about the need for Source, Scope and even Format since these may be obvious from accessing the version itself.

Anyway, we may then end up with only 4-6 fields, which seems to me to be more likely to be used.

So I guess that I am not in favour of describing "the ideal set of

information that could be available with a version" so much as the *minimal* set of information that could be useful, and a set of terms that will get us over the current confusion with "preprint", "postprint" etc.

I also think that the use cases showed that we can be document-centric rather than search-process-centric. That is, I think that thinking of use cases in the context of the document version rather than how the user got to that version is sufficient.

So my suggestion for the new year is that we focus on 1) the set of Version terms, and 2) a minimal set of metadata. Then we can apply these to the full set of use cases (maybe after having deleted the different search examples).

Cliff