

**Global Digital Format Registry Technical Working Group Meeting
November 6-7, 2006
Harvard University**

Attendees:

Stephen Abrams, Harvard University Library
Matthew Black, National Library of New Zealand
Andrew Boyko, Library of Congress
Adrian Brown, The National Archives,
Stephen Chapman, Harvard University Library
Brian Clark, OCLC
Laurent Duploux, Bibliotheque nationale de France
Adam Farquhar, British Library
Dale Flecker, Harvard University Library
David Giaretta, Digital Curation Centre
Rachel Gollub, Stanford University Libraries and Academic Information Resources
Leah Houser, OCLC
Larry Johnson, TethersEnd Consulting/NARA contractor
John Kunze, California Digital Library
John Mark Ockerbloom, University of Pennsylvania
Andreas Stanescu, OCLC
Ken Thibodeau, National Archives and Records Administration

Apologies from the notes taker for unintentionally missing the speaker's name on occasion.

Flecker: Welcome and brief introduction.

Abrams: Welcome, administrative details and introduction. We are looking for not formal signoff but a wider airing and a sense we are moving in the right direction.

Format Model

Section 2:

Giaretta: have you thought about extending the registry beyond formats?

Abrams: have considered this but the Mellon Grant prescribes a format registry.

Giaretta: this is sufficient and important work.

You don't have the information model in the format definition. Why? It is the domain of the FCS function. It is implicit. But this is an oversight and should be included. The FEF includes the FCS. You say they can be considered independently, but they are actually layered. What would it mean to conform to the TIFF format at an intermediate layer?

Ockerbloom: is there anything in this model that is used for other than relations? The number of levels is arbitrary – there are formats that may rely on fewer or more of the levels – the number of levels doesn't matter.

Abrams: the specific purpose of this model is to better define relationships.

Gollub: are you using the 4 level model from Jhove? What we're doing in Jhove can be expressed in a compact matter using this symbolism.

Giaretta: I think this is useful.

Ockerbloom: discusses his TOM encoding scheme.

Giaretta: we don't need a full functional definition of formats, but rather something that allows us to write software.

Be careful of your choice of formalism, because it dictates to some extent the concepts that you want to express.

Giaretta: When talking about formats, do you have document-like formats in mind only, or also scientific formats. Abrams: both. A file system is a format.

Ockerbloom: When you're talking about OAI Dublin core, do you add layers or a layer?
Abrams: The same 4 layers apply.

Abrams: maybe a little more thinking is needed on this (format model).

Section 3:

3.1 subtyping second bullet, Kunze: is this correct?

It is problematic that this document introduces a lot of new terminology. I can figure this out intuitively but that is not good to rely on intuition.

Johnson: It seems to rely on object oriented terminology.

Gollub: Would be helpful to have more examples.

Brown: we have not gotten this formal; this is useful.

Abrams: is the distinction between subtype and extension useful? Yes. There are some other relationship types, like all instances of a are also instances of b. We're missing the standard subtype relationship.

Abrams: we might want to explore whether to define relationships in such a way that the bi-directionality of the relationship is preserved.

Are the relationships defined broken, or do new relationships just need to be added?

Giaretta: Depends on what you want to get out of this.

Is the arrow in Section 3.2 extension pointing in the wrong direction in the example?
In this discussion, we are comparing the relationship between TIFF/DNG with ASCII/UTF-8.

Gollub: We need several examples in this Extension section.

Chapman: as a potential user of the registry, I need to have a clear sense of the inheritance descriptions, because I'm going to build on top of it. Need to make sure users of the registry understand this clearly so that people don't have to repeat all types of information that's already there. Therefore we can build on top of each other's work.

Johnson: once again we're jumping back and forth between graph theoretical terminologies and object oriented terminologies.

Flip the directionality in the diagrams. All of the relationships should have the same orientation.

Abrams: sounds like this needs more thought.

Stanescu: we need to add more examples.

Thibodeau: can a random subset of utf-8 be a subtype of utf-8? Yes. They do have a common abstract information model.

Caroline Arms draws a distinction between *can contain* and *must contain*. E.g. wave audio format, which can contain any type of sampled audio data, and broadcast wave which must contain PCM audio data. Also some formats can contain an instance of it, and some cannot.

Section, first sentence, 3.3 Containment, exchange the word 'subsets' with 'subsequences'.

ZIP files are a good example of containment.

Brown: it is a useful distinction to know that whether a format must be encoded within a container. There re some forms that can only exist within container.

Gollub: is that really a format if it cannot exist independently?

Abrams: must contain vs. can contain is a useful distinction. Wrapper vs. container distinction doesn't appear useful. Also distinguish whether the wrapper is required.

Ockerbloom: Needs to be *extends* or *contains* with respect to what.

Giaretta: it comes back the question, what does this information intend to be used for? For a zip file, how do you avoid an n-squared type relationship? Do you have to define ZIP as a relationship? What is the usefulness?

Brown: I'm not sure it is useful to define those types of relationships.

Giaretta: So this document declares: this is a set of interesting relationships between formats. It is not exhaustive.

Gollub: suggests a compromise position: this is a set of relationships that are technically useful.

Abrams: there is nothing that prevents someone from registering zip formats; whether someone wants to go through the effort, I don't know. The *must contain* relationship is a lot more useful.

Abrams: this document needs more thought and work. I will rework in light of your comments and repost to the wiki. I think Giaretta is saying we need use cases for the definitions.

Data Model Document

Introduction: this document presumes that the GDFR data model should be consistent with the PRONOM model. It is using PRONOM 4 as the baseline. (Brown: the current PRONOM data model is changing.) The substantive difference is dealing with format families.

Section 2.2.3, Pg 17

The format classification scheme is on the agenda for tomorrow. If we don't get to it we will discuss it on the wiki. There may be many classification schemes.

We re doing away with *family* as a descriptive element. We are using both abstract families and concrete formats.

Comments: this is where I was expecting the 4 layers to show up in this Taxonomy.

Abrams: it may end up like that for a set of formats, but not true of all formats. Do we accept this as is, or do we say that something about this is not right.

Brown: I think it would be useful to define this relationship more specifically, to tie the taxonomy with the 4 layers. Sometimes taxonomy does not line up with the abstract model, but it is still useful.

Brown: are there formats that show up in more than one family? SVG is part of the xml family and the SVG family. It would be more useful to be able to normalize data from the concrete to the abstract level. Also you should allow a format to be part of more than one family.

Stanescu: formats are being used at all of the 4 levels of the model. Bytestream, image, still and raster are all at the top level. But GIF is not shown as a relationship. How formats relate to each other on that stack is missing. XML can be used to encode structural relationships, or it can just be a format.

Discussion: There is no way you can create a plain hierarchy of relationships. Agreed, but you can create a pretty useful taxonomic structure. The taxonomy would be a view of a graceful relationship map. We like to print relationships out as if they were trees. We

could have different taxonomies depending on your point of view. We group them together because it is a useful way to discuss them because they have shared properties.

Abrams: but is the family mechanism repeatable? Yes. We probably need to go back to following the PRONOM model more closely. We want to assign more descriptive properties to the family type. We'll have to add format family back in.

Is GIF a different type of family than *raster still image*? Yes, but family is just another way of grouping together formats. Should classes be allowed to have full properties?
Gollub: Seems like something more generic that you could add text properties to would be the most useful.

Kunze: should we revisit the requirements for GDFR? That would help align our discussion. For example, Jhove is kind of silent partner here in terms of the problem of hierarchical relationships.

Brown, Stanescu: These classifications are useful in working out migration paths, to allow users to look at equivalencies between formats.

Abrams: Initially this is more of compilation or encyclopedia. If we highly type the information in the GDFR we could eventually evolve more actionable services.

Ocokerbloom: If we can get as far as getting a good robust way of describing formats and names that would be an extremely useful first step.

The most important thing is that we're talking about the same thing when we're talking about the leaves, the formats themselves. If someone wants to add yet more taxonomy that is fine.

The indentations in this diagram mean subclass. This is a comfortable way to discuss formats, although potentially everything is a graph.

Kunze: One of the obvious utilities of the GDFR is to look at the relatedness between formats.

Johnson: If you had the business use cases it would be a great way to vet what is in your data model.

<Session Break>

2.2.3 Format entity, continued

Abrams: we probably want to genericize the concept of format classification. Relationship should not be in there.

If updating the data model version is heavyweight, can you just update an enumeration? It will be a different version, whether it is minor or not. Consider a controlled vocabulary. Can we have closed vs. open enumerations, where open is based on an authority file?

Orientation element: this seems to be a weird division of text vs. binary.

Abrams: in general you would consider a text format to be more transparent than binary format. Some discussion. This could be misleading or deceiving. Suggestion: drop this as a characteristic. It is better dealt with as part of the classification scheme. Certainly this information could be found through other means.

Ockerlboom: one problem is having 3 different format specs because it can be encoded in at least three versions of utf. Do you have to assign an identifier to each possible encoding of the character format? Yes – Can we go with that for now but we'll have to fix it in the future? Perhaps a better solution is a relationship which is encoded-as, to handle this situation. We'll look at the latter as the solution. That means *Orientation* can be deleted.

Byte Order: Could be a relationship to an encoding format. Then, because UTF-8 has a byte order, some encodings would not have byte order because of that relationship. What we're saying is that there are some low level aspects that we would like to abstract away. So byte order can be part of the encoding format.

Internal signature: the value would be the regular expression. In PRONOM you may have more than one byte sequence that makes up a signature. This model needs to allow for one or the other or both. You also define a region, it's not just an offset, where you look for the signature. This should be added to the model. Consider using content type or mime type as the *external signature*.

Is metadata about the object within a container considered internal or external to the format? This should be considered external. We should generalize this case that you can look for metadata about the object to determine its format. An internal signature in a zip file could describe the zip file itself, in the case of an open office file.

Do we need to worry about describing whether there is a file inside the container that describes the files it contains. This is a rule for parsing. How do you specify this? Is this needed? Depends on the use cases. This would work for format identification and verification. Agreed to save this problem for later.

Reference file: how does this differ from documentation? Documentation is human-interpretable. This is an example of the format being described. More usefully this could be a set of conformance files.

Grammar and Assessment

Grammar - enables future value-added services that could be machine-level. Not sure if these are the right grammar types. EAST may be better used to describe an actual instantiation of a format. Each grammar should be a format in its own right. It should provide enough information about the grammar to aid in the interpretation.

Is human-readable BNF a format? No, we think it is documentation. Machine-actionable BNF is a format.

Giaretta: the line between human and machine-readable documentation seems arbitrary. He suggests not distinguishing between the two. Would you then want some indication of whether the documentation is intended for a computer? Not sure.

Assessment – LC has an assessment method with a structured form. Brown: we're going to be adding risk assessment information to PRONOM. We decided to support a variety of types of assessments since there is not a prescriptive format. Brown: we've been trying to base the risk assessment on information which is already in PRONOM. There is a mechanism to recalculate the risks from that data. There is information on how to interpret the information in those fields. For now it is easier to make the PRONOM assessment a type. This makes the model more extensible over time.

Withdrawal date – you may have items that are outdated and still supported. You may know a withdrawal date sometime in the future.

Rights – “in the public domain” is not a concept in every jurisdiction. Need an IPR entity for the U.S., etc. Add ‘public domain’ to the IPR Type domain.

Question about license type: the key thing is actually the license form. Adding ‘License’, which is of type Document. Removing ‘license type’. Under ‘license form’ remove ‘other or unknown form’ and replace by “not one of the above”. Add ‘Public domain’ as a type of IPR. Special license arrangements on an institution level are out of scope.

Do we want to come up with a few easy checkmarks to specify commercial vs. non-commercial, for example? Perhaps we could look at Creative Commons for this.

Needs to be a license form which is ‘proprietary’. Do you also want to specify license domain, such as unrestricted. Also, are there other types of documents that you want to link to this, such as patent documentation. Patent Number could be added to the Identifier Type.

Abrams: so we need to flesh this out a little bit, then.

Back to Format...

Status is covering several different things. We need to be more precise, or leave it out.

Brown: regarding assessments, in PRONOM we're going to make it very clear what is objective and subjective, and disclaimers, etc. This is worth thinking about further.

In GDFR we are strictly limiting the assessments to risk assessments.

Is the system identifier unique across all nodes? No, this makes life more difficult. This is persistent to the node, but not unique across all nodes.

Need to make versioning more obvious in the Base entity properties. You'll be able to request the versions. You'll be able to ask for a specific version using a unique URI.

November 7

GDFR analysis model & use cases

Stanescu reviews the format registry wiki at www.formatregistry.com.

Section 3, Requirements

Change “the set of registries” to “the set of nodes”.

Kunze: Is it possible for registries to maintain info that is not replicated cross nodes, i.e. local records.

Stanescu: That would be possible, but it would be better to have the records distributed to a limited group of other nodes to protect the data.

Ockerbloom: If a record goes to another node, is it replicated in full? Yes. Are the publishers of this information OK with that?

Abrams: There may be some sort of escrowing mechanism which refers to documents. A particular institution can register a ‘local holding’. The document would then be managed in an external manner (i.e. external system) so that it is not replicated across the nodes.

3.1.2 Data Requirements

Ockerbloom: is ‘vetted’ a binary state or does it include agent information?

Abrams: We are assuming there would be a single GDFR-managed process. All data elements being managed within the network would be tagged. That would include the agent that submitted the information. We need to consider how we will keep others from setting the ‘vetted’ tag.

Thibodeau: as it stands, is vetting outside the scope of the software to be developed?

Stanescu: We will start talking about the vetting process in the next few months. The form that this takes must be discussed. We may use an existing process. I view the vetting process as a different ‘component’ than the registry.

Ockerbloom: does the model support both the vetted and unvetted version of the format simultaneously?

Abrams: it's not quite clear what granularity is required for tracking the elements within the record. Comment: it seems that the status is at the record level.

Stanescu: perhaps parts of the record are vetted, and some parts are unvetted. For example, perhaps some of the assessments are vetted, and some are not. Does the record go through a lifecycle of unvetted to vetted? Seems like there should be one blessed version of the record.

Johnson: You will need a model of authority in terms of who authorized the record. Each action to the record should be tied to an authority.

3.2.2 Versioning

Stanescu: the only 2 things specific to the GDFR registry are the data model and the distribution. (The rest of the functionality is general to any registry.)

Gollub: please add this clarification to this document.

Kunze: are you thinking beyond GDFR to other types of registries?

Stanescu: Yes. For example, the authority information is just another registry collection.

Brown: change *impervious* to *resistant*, under section 3.2.3 Synchronization of Nodes.

3.2.4 Registry Configuration Policies

Thibodeau: are you assuming that each node does configuration management? Yes, this is a network of cooperating nodes. But you will be able to have administrators 'remotely' administrating nodes.

Figure 1: Registry domain model

Stanescu: the relationships between the collections of objects/records in the registry also constitute a collection. The format record points to other objects – it does not encapsulate them. It is something like a mashup. It will look as if it is contained to the user.

Ockerbloom: how does this interact with the whole vetting process?

Stanescu: I think that only the format record can be vetted.

Gollub: we need info to clearly mark what is vetted and what is not. We would need a clear definition of what parts of the record are vetted.

Brown: in vetting a record, that includes reviewing everything that is linked to that record, including the IPR and various other things it points to.

Stanescu: I think that makes sense that it is in its entirety.

Ockerbloom: how can we say we vetted the IP rights except in relation/context with the particular format? So it is effectively contained in this case.

UML Model diagram

Johnson: The UML diagram needs to indicate cardinality. It also needs to clarify what is contained, and what is an independent object. The class model needs to capture the semantics of your space – we need cardinality in order to document this.

Stanesco: I will put in the cardinality. This diagram will also need to reflect changes in the datamodel. Needs to document the constraints in some language. This is a domain mode, not a class diagram.

Stanesco: There is a third relationship, which is ‘encoded by’. Another subclass should be added.

Johnson: which entity in here corresponds to the record? Just the Format box. We need to limit our vocabulary to the entities in this document. The Format box should be called Format Record. Don’t use the word ‘record’ unless it’s in the document.

Johnson: This diagram should be expanded into a complete class diagram.

Stanesco: I will do that but it is too soon now.

Brown: do you mean ‘generates’, or something more? Perhaps ‘process’ is better. Generate is too specific.

Johnson: it may be preferable to organize this document more clearly in terms of the 5 engineering viewpoints.

Stanesco: I’m going to take out the networking view and put it in another document.

Johnson: GDFR is larger than just the software being developed; it is also the ways in which the software is being used for some business purpose. Something that even has an ROI associated with it. You may find some blind spots as a result of doing this (i.e. thinking more about the business use cases.).

Stanesco: I would like to do this; I will do this within the next month or so.

Thibodeau: Why does only hardware have a relationship to file?

Stanesco: oversight :)

Section 5 Use Case Model

5.1 Actors

Fix typo: “an extension o the publish user”. Also the line from registry editor to export should be removed.

Gollub: can you export previous versions? No, the export use case only exports current versions. (You can get previous versions of a record individually.)

Chapman: are you assuming that this registry will not have a delete function? Yes, you update the record with a delete status but the record is not deleted.

Brown: under any circumstances you will not allow deletion? If I had to, I could delete it somehow?

Chapman: you should consider that a superuser to be able to do this. You may need to do this for legal reasons.

Abrams: we want to add a purge delete case for completeness as a rights restricted administrative function. We may also need user case for adding new nodes.

Stanescu: I will add a specialized actor for this.

5.2 Use Cases

Walker: is the concept of merging format records covered here? In the case that 2 different nodes created a record for the same format?

Stanescu: The concept is not directly built in. For right now, this would just be an extension of update.

What if the synchronization of formats notices that the records are out of sync? That requires human intervention. It is an exceptional condition of the synchronize. It is recoverable by a simple update but there needs to be a decision maker. This might be the super user role. How do we know which version of the record is accurate? We don't necessarily have a master node. We can discuss this further this afternoon when we talk about the network and synchronization.

Let's also talk about disaster recovery this afternoon.

What do we think the lifetime of this software is? Forever?

Abrams: 10 years?

Hopefully the information content will last much longer than that.

Add and Create.

Gollub: I'm not sure why create and add are separate, even though I read the doc.

Stanescu: it is a factory pattern in which the factory instantiates the object. The identifiers are created as part of the object creation. It also facilitates user interface creation since the interface gets an object it can just fill out.

Gollub: this has a large number of schemes for identifiers. Are you going to recommend one? There will be a URI for each format.

Ockerbloom: perhaps you should show 'create an identifier' in the use case model.

Stanescu: institutions can act as mirrors, which lowers the barrier of entry to this network. Every source node is a mirror node.

Abrams: you may need a vetting process to determine who gets to be a source node.

Stanescu: I need to consider how a node gets added to the network. If you're a source you're a source – you are not precluded from adding/modifying any of the records. There are no restrictions on who can become a source of any record.

Stanescu: an institution can use another institution's service to be a source.

Gollub: How do we keep duplicate formats from being added to the registry? We've been working on software to compare titles at a word map level. Perhaps you could use this.

Import use case:

Import is really a batch create and add. The use case should say 'the record should be properly formatted', as a precondition. The use case should also say 'does not contain GDFR identifiers'.

Export use case:

Thibodeau: if there is a record that is vetted, and it replaces the unvetted record, the user can only get the vetted record? Yes, we're only going to keep one version of the record.

It seems that import and export are not symmetric. The quibble is with the naming that infers that they are symmetric when they are not related to each other.

Brown: I'm thinking about what it means for an existing system to become a GDFR node. It is not clear to me yet. It seems desirable to have many implementations of format registries. This is an open issue in terms of the relationship between PRONOM and GDFR.

Synchronize:

There should be 2 use cases for sync: harvest and synchronize. One of the responsibilities of being a source node is that everyone will come to you eventually for that record. A new node, any node, always comes to the source node for a record.

Gollub: are you talking to the folks at LOCKSS about this?

Stanesco: yes.

Authenticating the user:

Stanesco: Liberty Alliance or Shibboleth is what I'm thinking.

Thibodeau: where would you fit subscriptions (subscribe to updates)?

Stanesco: It would be in the distribution model. Public users can subscribe. We need to add subscribe/notify as a use case.

Some business use cases/value added services:

Where to find documentation about grammar – will do this now

Is a format well-formed

Format validation – source of grammar

Format identification

Format aliasing

<lunch>

6. Component Architecture

Distribution model

Stanescu proposes that we use LOCKSS as the distribution mechanism for the registry information. Discusses paper on the use of LOCKSS in a preservation system. We would configure the LOCKSS distribution for the nodes in the network. We can use the implementation and model as they currently stand. We do need to do something about access control. We could implement a façade to do this on behalf of an external registry.

Another problem: what happens if the registry goes out of business entirely? Another node needs to become the source for those records as a policy change.

3rd problem: the node is temporarily down, so the format record cannot be modified.

Chapman: on 2nd issue, is it a problem for a node to take over another node's records, in terms of the reputation of the other node's information?

Farquhar: Is the LOCKSS solution solving threat problems that aren't important to us? Also, we don't view the registry as a preservation system itself (it is not a digital archive.) Also argues that LOCKSS is not well-established software and is not where we want to take on risk in this implementation.

Farquhar: do we actually need a peer to peer network?

Stanescu: we don't want a single point of entry because some company has to make a commitment long-term to it, and also because of availability. It is desirable to distribute the point of entry.

Farquhar: if nobody cares to support it any more, there are no more updates.

Discussion: we're assuming that the cost to maintain this service is very inexpensive.

Abrams: I think another institution would be agreeable to taking on hosting of the root node, unless no one sees value to this any more. The governing group of institutions may need to agree to become the root node.

Frequency of update on GDFR may be a few updates per day.

The question is whether there is a single or multiple root is not tied up with the use of LOCKSS.

Kunze & Farquhar: other alternatives to this architecture would be (1) a zip file, (2) dspace with additions.

Gollub: I think the distributed model lets us get way from the ownership idea, which is positive.

Chapman: Having to run our own system is a bigger barrier of entry. With this model there's little to nothing that we have to build locally.

Gollub: my team is building OAI support for LOCKSS right now. Delivery is a few months away.

Ockerbloom: what is the granularity of the records that we are going to transfer? OAI supports single objects that are not related. We will need to think about the granularity issues.

Gollub: would it be helpful to have one of the LOCKSS people around to help make the decision?

Abrams: how important is the validity and assurance of the data? Some sort of auditing function to assure consistency across the mirrors is important. Agreement on this.

Farquhar: this is a small amount of short-lived data.

Abrams: doesn't agree on short-lived point. Over the next year the effort should be invested in data.

Abrams: I want to be assured that my data is up-to-date, but not at the cost of considerable expense.

Stanescu: One issue with LOCKSS is the frequency of sync at 2 weeks. The LOCKSS staff have stated this is not an issue to change.

Kunze: my institution would prefer not to use LOCKSS for this – its complex and it would create a dependency on LOCKSS.

Abrams: we could choose to make sync optional.

Mirror diagram:

Stanescu: I think we are walking way from this idea of having different (multiple) data sources.

Section 7, Service interfaces

Stanescu reviews the services available from the WSA interface. Shows demo of the prototype at the collection of collection and at the registry level. Discusses the GDFR node diagram drawn as the IWSA framework.

Thoughts on XML database: the protocols are more interesting than the database implementation, since the database should be relatively small.

Gollub: is performance with db xml is pretty good? Have you load tested it?

Abrams: at Harvard we have 3 million identifiers with thousands of hits per hour – it has been fine. We need to make sure that the xml layer on top of this is as solid as the database.

Abrams: I don't see any reason not to go ahead with this as the initial prototype.

Policy Considerations

Thibodeau: Larry Johnson's role is to come up with a governance model for this work, as a separate project. There are dependencies between the architecture and the governance. Hopefully we can keep the governance as independently of the other layers as possible.

This is a parallel effort. The outcome is to simply to define the process that will define the governance.

We need to revive the higher level use cases and take a look at them to ensure that the stake holder's needs are being met. Harvard's obligation is to hand over responsibility to some other institution that remains to be identified. We have undertaken maintenance activities for 2 years if no maintenance organization steps forward.

Other policy considerations:

Institutions could have a number of relationships to GDFR: simple source, mirror node, involved in vetting, a consumer. What is the motivation for an institution to take an active role in any of these capacities? So far the considerations are negative. We would need access to the history records. What is the business case for my stepping up to be a source node?

Abrams: we would do so using locally developed expertise and would share this as another mechanism to continue the same activities we've always undertaken.

Johnson: what would be the role of Microsoft, Adobe and others?

Abrams: it would be in their own self-interest to make this information available.

Farquhar: I don't think there is any barrier for them doing so.

Thibodeau: your service model leaves open the possibility of an organization using the registry to create other services that are not preservation-related.

Chapman: in some cases you are talking about archiving, but this is not narrowly constrained for simply that purpose.

Johnson: is it the intention to have a complete formal description of these formats?

Abrams: no, if you have that data you can put it in, but it's optional.

Any Other Topics?

Giaretta: in the format description, in the grammar, regarding the formalism of the grammar describing the format, how will people know how to understand those formal descriptions?

Abrams: Those descriptions would be references again to formats themselves. So if the grammar is typed we would have a tiny representation network into a format.

Giaretta: when you move away from things that are simply rendered, then a grammar is not sufficient.

Abrams: my assumption is that we want to take as expansive a definition of format as possible.

Johnson: we need to work out how the formalism of defining the format is usefully connected into the rest of this work.

Abrams: I need to look at this more closely.

Farquhar: So a question is, what are the ambitions of GDFR in the space where people create their own formats, such as a comma-delimited file.

Abrams: we could do something like that in a human-actionable way, but not in a machine-actionable way. It is the intent to support this scenario.

Houser: the problem of what formats are of interest is based on the community working with it.

Gollub: we found that several formats could be classed together with human-readable data; with others we needed more formalism.

Carl F. and Caroline Arms have a workflow problem because they are constructing things by hand, since they don't have the advantage of inheritance. They want to make sure that 'relationship' has enough functionality to take the pain out of that process.

Abrams: I don't want to preclude communities like scientific data set communities from this work.

Thibodeau: so what type of representation information is out of scope? Data dictionaries and code books are in scope.

Giaretta: I think conflating structure with semantics makes this problem more difficult.

Farquhar: this effort appears to be closer to the structure side.

Brown: we have tended to focus on the structural end of things. Not get too far into the muddy waters of representation information. There hasn't been a formal test to apply to determine what is in scope in PRONOM; similarly no formal test as to what is a format. It would be useful for GDFR to get to the bottom of this.

Brown: we added the ability to add temporary identifiers because you might put something in and want to split it into 3 things later. You might want to consider records with provisional status when they are being worked on.

Stanescu: Perhaps this is a workflow issue – the state of the record changes instead of the record identifier. Perhaps you could put restrictions on its distribution, e.g. make it a local record.

Stanescu: maybe one of the format relationships is 'successor'.

Next Steps

Abrams: A lot of revision within the docs we've reviewed. Attempt to continue this level of discussion on the wiki on the discussion page for each document.

Q: Is it possible to strike out closed issues to make clear what is still hot topic?

Clark: we can delete comments themselves. We need to know what is still a hot topic. Perhaps we can mark closed topics – that seems to make sense.

You can configure the watch list on the wiki page so you are notified of changes. The discussion is with the article page, not with the documents, so the conversation doesn't apply to a version of a document. We'll probably create a new discussion page when we update a version of the document. We will keep old discussion pages.

There will be another iteration of these documents, hopefully capturing the consensus opinions from this meeting. We would like to publish publically early on in the next calendar year.

Do we want to continue convening this group? Matthew has worked out a call schedule :) Potentially we could have a follow-on face-to-face meeting, especially for items we did not get a lot of consensus on. Perhaps the West Coast next time. Regarding discussion on the wiki, you can watch for updates of the documents, and also new conversation.

Matthew: the wiki doesn't appear to have a 'watch this page' feature.

Clark: I will follow up on this.

Giaretta: it is desirable to have the changes digested and sent periodically. We can also have scheduled meetings on the wiki on specific topics.