LarryStone

I'll consider renaming "Quality" to "Confidence", although gives identification methods no basis for comparison. The level values are a small number of enumerated points. I don't believe the level should be real, since that identification Is your reason behind a real number to allow later additions of constants in between the current ones? I agree that's worthwhile, but we can possibly do that with an integer, just space out the initial values by 100 or so (remember BASIC line numbers?).

Robert Tansley 31-Jul-07

You've incorrectly assumed the original intent of the 'internal' flag (you never asked?). It was originally intended to flag formats that were DSpace-specific and for DSpace's internal use. Hence it was intended to describe the format. That the UI used this as a hint was a side effect; further, others started to misuse the flag and/or use Bundles as the hint instead. I did post this question to a list some months ago. "License", in particular, does not describe the format in any useful way since it is erroneously applied to RDF/XML, plain text, and HTML files. I don't believe BitstreamFormat needs an "internal" bit at all; should there be two separate BitstreamFormat's for the "RDF/XML" format, one with the internal bit set and one without? If anything, that bit belongs on the Bitstream object. Then it could usefully identify e.g. Bitstreams used as logo images in the UIs. --LarryStone

You still haven't addressed the .avi use case, where the encodings of streams needs to be found (never responded to my comment on 'About Data Formats') Assuming 1-1 Bitstream - Bitstream Format is naive and not sufficient for numerous use cases (video, multimedia presos, XML, etc.) You can quite reasonably choose not to address this in the first pass but saying 'there are no use cases' is incorrect.I looked into AVI; it's a Microsoft-proprietary wrapper format that is specialized to contain mostly video and audio streams (of various formats) embedded in it. Logically it's similar to a restriction on a container format like Zip.

I've observed container formats can mostly be classified as one of these two:

1. General-purpose container like Unix "tar" or Zip, member files of any formats with no restrictions or expectations.
2. Specialized container for a particular application with structure and restrictions imposed on contents, e.g. JAR, WAR, IMSCP, AVI.

In the first case, it's easy to argue that the container itself is not significant and if you want the contents of the container preserved in the archive, unpack the container (and preserve its existence as logical relationships in the metadata of the individual Bitstreams). This is how we handled IMSCP for OCW. Alternately, treat the container as one Bitstream, it has a format.. but the contents of the container are invisible to the archive.

In the second case, the container is a logical unit; it is only meaningful to applications which expect and understand it. It should be treated as a single file by the archive. Any preservation tools that handle it will have to know how to unpack and repack it, anyway.

So what I mean by use case is not "are there files that could be described as a hierarchy of formats", but "is there a purpose to describing the format of a Bitstream as a hierarchy of formats" in the archive, what would DSpace do with that knowledge?

Even the GDFR format model (the most sophisticated one around) gives up on describing instances of container formats. It describes a container format as possibly containing or necessarily containing such-and-such other formats, but all instances of the container are described by that same format entry. --LarryStone

The set of 'quality' levels seems rather arbitrary. A real no. between 0 + 1 might work better, perhaps with 'user' as a special value. 'Confidence' would be a better term. That USER 'overrides' an automated identification seems problematic -- end users are not necessarily the experts on exact format identification.Is your reason behind a real number to allow later additions of constants in between the current ones? I agree that's worthwhile, but we can do that with an integer, just space out the initial values by 100 or so (remember BASIC line numbers?). I don't believe the level should be real, since that gives identification methods no basis for comparison. The level values are a small number of enumerated points.

I'll consider renaming "Quality" to "Confidence", although "confidence" has an unfortunate alternate meaning wrt. establishing veracity.
I'll clarify this in the documentation, too. --LarryStone