Introduction to SIARD 2.2 RFC.

The SIARD file format version 2.2 Request for Comments is based on SIARD 2.1.1 and is strictly focused on what is needed to ensure scalability, esp. handling large objects outside the SIARD file according to SQL:2008, chapter 9 SQL/MED. SIARD 2.2 should be fully backward compatible with SIARD 2.1.1.

All other requests for change, be it additions, replacements or removals to the SIARD File format specification have been ignored and postponed to the next revision, which we expect will come soon after this, led by the DILCIS Board.

We urge all current and coming users of SIARD to give their comments to this Request for Commons version 2.2 of SIARD. And those being part of the process of creating this RFC should hold themselves back from commenting again.

Having received these comments the DILCIS Board will modify the specifications and issue a new version of SIARD named SIARD 2.2.

In the following sections we have some specific questions for you:

Should the folder *header* be repeated or not in segments?

The folder *header* contains the files metadata.xml, metadata.xsd and the folder siardversion, containing a folder with the version. Additional files, such as style sheets, are permitted (see P_4.2-5).

What should be done with the header folder when the SIARD file is segmented?

Should it be repeated in every segment or only exist in the first segment?

And if the folder is repeated, should all its content (files and subfolders) also be repeated?

In this RFC the header folder is not repeated when LOBs are store outside the SIARD file (section 7.1, 7.2)

In this RFC the header folder is repeated with all its content (files and subfolders) when the SIARD file is segmented due to large tables (section 8.5 and 8.6).

The answer to these questions may depend on one's preference for unambiguity compared to self-sufficiency i.e. avoiding repeating information possibly increasing the risk for inconsistency (once only please) compared to possibly being able to process each segment by itself.

We have hope for a clear answer, so we can use the same principle for both cases in the final version 2.2

Question: Should the folder *header* be repeated in all segments?

Question: If yes, should all its content (files and subfolders) also be repeated?

A manifest for a SIARD file with data outside it?

A SIARD file without data outside itself

The SIARD file format has never had any manifest file (such as a file containing filename, path, and checksum for all files in a folder hierarchy).

A SIARD file is assumed to be archived as part of an information package (see section 1.2.3 of the SIARD Format specification) which is expected to have a manifest file to ensure fixity.

Should it somehow be necessary to know the amount of files in a database in SIARD format one will have to parse metadata.xml to figure out how many table[n].xml, table[n].xsd, and record[o].bin it should contain in the folders under *content*. Normally the program creating a SIARD file will take care of this consistency. If needed a checksum of the SIARD file can be created by the program before placing it in an information package.

A SIARD file with data outside itself

When a SIARD file refers to files outside the SIARD file itself (as in the case of LOBs stored outside the SIARD file or a segmented SIARD file due to size) the question arises if that makes a manifest file necessary or not. Normally the program creating a SIARD file with data outside should take care of this consistency. Nevertheless, there may be a higher risk of losing data outside the SIARD file, especially if manual operations (people) are involved.

Question: Is a manifest for a SIARD file with data outside necessary?
Question: If not, are you against adding one anyway?

Placement of the manifest file

Provided that the answer to the question "Is a manifest for a SIARD file with data outside necessary?" is a sufficiently clear "yes" then we need to decide where to place such a manifest file
The placement could be inside the header folder, or it could be next to the SIARD file.
A benefit of having the manifest file inside the header is that it is not easily lost.

Question: Where should the manifest file be placed, inside the header folder or next to the SIARD file?

Format of the manifest file

An even larger question is the format of the manifest file. In itself it is somewhat problematic that digital archiving does not have one common manifest file format, but that we seem to use several slightly similar formats.
Below is a list of formats in decreasing complexity, and we urge you to tell us which one you prefer.

a) Oxford Common File Layout Specification, https://ocfl.io/1.0/spec/

b) RFC 8493 - The BagIt File Packaging Format (V1.0), https://tools.ietf.org/html/rfc8493

c) Checkm: a checksum-based manifest format, http://jkunze.github.io/checkmspec.html

d) Hashdeep Fileformat, https://github.com/jessek/hashdeep/blob/master/FILEFORMAT

e) MD5Sum Format,
https://www.gnu.org/software/coreutils/manual/coreutils.html#md5sum-invocation