4th International Digital Curation Conference
December 2008

DCC DIFFUSE Standards Frameworks: A Standards Path through the Curation Lifecycle

Sarah Higgins,
Standards Advisor,
Digital Curation Centre

October 2008

Abstract

DCC DIFFUSE Standards Frameworks aims to offer domain specific advice on standards relevant to digital preservation and curation, to help curators identify which standards they should be using and where they can be appropriately implemented, to ensure authoritative digital material. The Project uses the DCC Curation Lifecycle Model and Web 2.0 technology, to visually present standards frameworks for a number of disciplines. The Digital Curation Centre (DCC) is actively working with a different relevant organisations to present searchable frameworks of standards, for a number of domains. These include digital repositories, records management, the geo-information sector, archives and the museum sector. Other domains, such as e-science, will shortly be investigated.

The 4th International Digital Curation Conference taking place in Edinburgh, Scotland over 1-3 December 2008 will address the theme *Radical Sharing: Transforming Science?*.

Introduction

Effective long-term curation and preservation of digital information relies on the implementation of appropriate standards and technologies which support curation processes over the entire lifecycle of digital material. With hundreds of standards, and multiple versions of these to choose from, selecting those suitable for curation and preservation actions can be a daunting task.

DCC DIFFUSE Standards Frameworks[1] offers domain specific advice on relevant standards. The DCC Curation Lifecycle Model (Higgins, 2008) is used to contextualise standards and visually present searchable frameworks of these. This helps digital curators identify which standards they should be using, and where they can be appropriately implemented. The use of Web 2.0 technology encourages community engagement and ensures information regarding standards usage is maintained. Meanwhile the Project is actively working with organisations from a number of disciplines to identify and develop appropriate frameworks.

Standards in Digital Curation

Information technology standards facilitate the implementation of solutions for creating and storing digital material, and their subsequent access, use and reuse. Many standards are generic, and offer functionality which can be used across different disciplines, to enable curation aims to be achieved. These include file formats, reference models such as OAIS[2], persistent identifier standards and standards designed to enable remote access, deposit and authentication. Other standards are discipline specific, and have been developed for a particular purpose which is not widely applicable. In particular, metadata standards, authority files and XML compliant mark-up languages are often very specific to the material being described, and represent the result of intra-disciplinary collaborations. Examples of these are widespread and include; the metadata structure standard ISAD(G) (General International Standard Archival Description)[3] designed for describing archives, which can be marked-up in EAD (Encoded Archival Description)[4] and mark-up languages such as Chemical Markup Language[5] and MathML[6].

Implementors combine both types of standard, to develop frameworks which can be used to effectively manage their digital information. These frameworks can also be the result of disciplinary collaborations, with domains sharing the problem of identifying sets of standards which can achieve a community aim.

Implementing standards frameworks can have multiple community benefits, including: encouraging the achievement of community objectives through consistent and increased participation; sharing of resources, procedures, architectures, metadata profiles and access terminologies; and interoperability of hardware, software and data. Developing and sharing effective standards frameworks can increase business

[1] http://www.dcc.ac.uk/diffuse/
[2] ISO 14721:2003 Space data and information transfer systems — Open archival information system — Reference model
[3] http://www.ica.org/en/node/30000
[4] http://www.loc.gov/ead/
[5] http://cml.sourceforge.net/
[6] http://www.w3.org/Math/

effectiveness through efficiency savings and ensuring legislative compliance. At the same time the implementation of sustainable and viable systems, with effective workflows, can be undertaken with reduced organisational design work. For some disciplines frameworks are *de facto*, relying on community agreement for their application, for example the Joint Information Systems Committee (JISC) Standards Catalogue[7], which details recommended standards for the projects they fund. Other frameworks, such as the UK Government's *e-Government Interoperability Framework (eGIF)*[8] are mandated to ensure interoperability across a sector.

The benefits of standards frameworks to ensure consistency of approach, and consequent interoperability and collaboration, have been explored by the UKOLN Interoperability Focus[9]. Interoperability with the cultural and heritage sector is discussed by Gill and Miller (2002) and reports, from orchestrated meetings with the sector, saw a willingness to collaborate on defining standards frameworks (Miller et al, 2001). The benefits of open standards to the digital libraries community, to avoid vendor lock-in, and effective collaborative implementations have also been examined (Dunning et al, 2005).

Standards Frameworks and the Curation Lifecycle

The continuity of digital material is best assured when the lifecycle approach to their management is undertaken. The benefits of this approach for archiving digital objects was discussed by Hodge (2000), and for curation of digital information by Pennock (2007). The DCC is committed to promoting the lifecycle management of digital assets, and has developed the DCC Curation Lifecycle Model (figure 1) to facilitate planning. This lifecycle approach to curation needs to be underpinned by the implementation of appropriate standards and technologies. The Model can facilitate planning frameworks of these, which ensure support for all parts of the lifecycle.

Standards frameworks intended to support the curation lifecycle, should ensure that the recommended technologies maintain the authority of digital material, as defined by ISO 15489. Authenticity (the material is what it puports to be) is maintained through: access controls; appropriate metadata; consistent use of persistent identifiers; and bitstream calculations such as checksums to ensure data has not been corrupted or tampered with. Reliability (the contents can be trusted) is ensured through the maintenance of complete, organised and accessible material. Integrity (the material is complete and unaltered) relies on protection by authority control. Usability (the material can be located, retrieved, presented and interpreted) is maintained through: the implementation of systems appropriate to the *business* aim; inclusion of a comprehensive range of material for contextual understanding; and systematic management of material throughout the lifecycle. Additionally, standards frameworks for curation will ideally support interoperability, maximise accessibility, avoid vendor lock-in, provide architectural integrity, and help to ensure long-term preservation.

DCC DIFFUSE Standards Frameworks

In some domains, the benefits of standards are well understood, and

[7] http://standards.jisc.ac.uk/catalogue/Home.phtml
[8] http://www.govtalk.gov.uk/schemasstandards/egif.asp
[9] http://www.ukoln.ac.uk/interop-focus/

comprehensive frameworks have been developed, documented and made accessible for potential adoptors. DCC DIFFUSE Standards Frameworks has started to capture these frameworks to make them further accessible. The Project graphicallys explain the way in which the standards included in a framework can be concurrently implemented to achieve curation aims. At the same time the Project can act as a depository for organisations, consortiums and projects, enabling them to document the frameworks they have developed in a consistent manner, manage them in one location and advertise them to those seeking curation solutions.

The resource consists of a browsable database of standards relevant to digital curation and preservation. Users can opt to browse by choosing a relevant Framework. The DCC Curation Lifecycle Model then offers a graphical searching tool, indicating the appropriate stage for implementation of the standards documented within the Framework. These contextualisations will help users to readily identify which of the many standards included in the database are appropriate to their own situation. They can identify which are designed to support the curation actions they wish to plan, aiding informed choices regarding implementation. It also enables gaps in the curation planning process to be identified and areas where additional standards need to be considered, or even developed. The database also offers a number of other browsing options: by title; by the technical function they support; and the organisation responsible for their development.

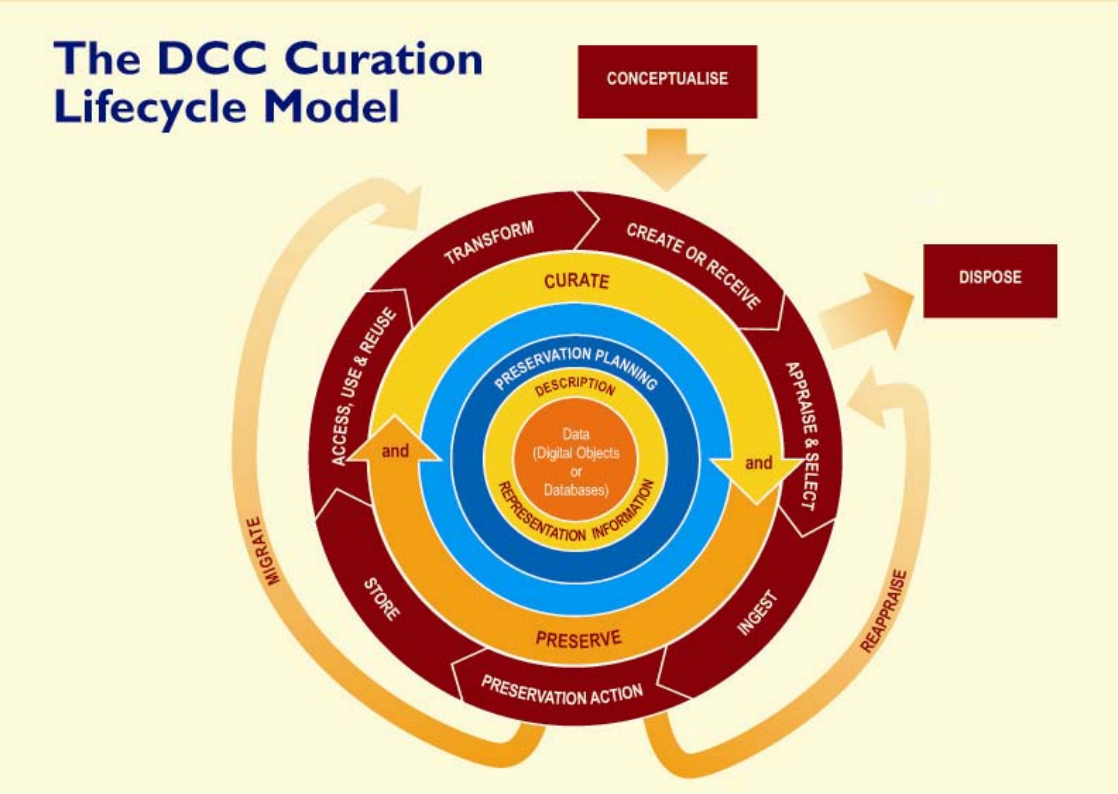

Figure 1 The graphical component of the DCC Curation Lifecycle Model (Higgins, 2008)

Information regarding individual standards is being described using a profile of

the Standards Metadata Element Set, v3.0[10], which was specifically developed for documenting standards by the ANSI (American National Standards Institute) hosted Standards Registry Committee[11]. All standards are classified according to the frameworks they are included in, the lifecycle function they support and their technical function. Users are able to identify: previous versions of a standard; those standards which are referenced within a standard and need to be used in conjunction with it; and those which have been created by the same body. Descriptions link to: standards documentation; information concerning sponsoring bodies; and further useful documentation concerning a standard such as implementation guidelines, xml schema or best practice guidelines.

The English language version of the Website Wikipedia[12] is being used to manage standards descriptions and encourage community participation. Early test descriptions created for the Project, showed that this was the first resource consulted when researching data for the fields documenting both functions and usage of standards – particularly as the actual standards documentation is not always readily available. This, and the enormous task of keeping the descriptive information up-to-date with limited staffing and budget, led to the decision to encourage the community to undertake the maintanence of this information. The possibility of a custom built DCC DIFFUSE wiki for this purpose was considered, but it was decided that this was unlikely to achieve the community buy-in possible with the high profile and apparantly stable Wikipedia. DCC DIFFUSE will create Wikipedia entries where none currently exist, keep a weather eye on existing descriptions, make occassional corrections and updates, and encourage users, and collaborators, to do likewise. Research is currently being undertaken concerning the feasibility of linking to a particular "DCC endorsed" version of a Wikipedia entry, and the possibility of formatting the standards pages within Wikipedia consistently, for future harvesting into the DCC resource.

Contributions to DCC DIFFUSE Standards Frameworks

The Project is currently working with both the digital repositories community and the records management community to capture the frameworks identified by the Driver Project (Foulonneau and Francis, 2007) and MoReq2 (Serco Consulting, 2008) respectively. It is hoped that these 2 frameworks will be completely documented by the end of December 2008. Collaborative work will shortly begin with both the Open Geospatial Consortium (OGC)[13] and the UK Society of Archivists Data Standards Group, for documentation during spring 2009. The former is likely to undertake data entry for the Project for the standards they develop and recommend. The latter have been developing a framework, which is currently the focus of a series of articles in their member's newsletter, *Arc*. This framework will be further developed, in conjunction with the DCC, for presentation in DCC DIFFUSE. It is hoped that it will also be accessible from the Society of Archivists website[14], ensuring maximum

[10] http://publicaa.ansi.org/sites/apdl/Documents/Other%20Services/Standards%20Registry%20Committee/Standards%20Reg%20Metadata%20Def%20v3.0.pdf
[11] This Committee reported in March 2003 – minutes and the metadata set can be found at: http://www.ansi.org/internet_resources/standards_registry_committee/stdsreg.aspx?menuid=12
[12] http://en.wikipedia.org/wiki/Main_Page
[13] http://www.opengeospatial.org/
[14] http://www.archives.org.uk/

exposure to relevant adopters. Meanwhile discussions are underway to work with representatives of the web archiving, the museums, and the particle physics communities, to identify standards frameworks for their disciplines, document them and make them accessible. Plans for cooperation with the e-science sector, through the e-science liaison function of the DCC, are also being developed.

Possible Further Growth

Initially the DCC approached organisations asking them to participate in DCC DIFFUSE. Latterly the situation is reversed so that a demand for the services it offers and potential for growth, has been identified. In addition to the extensions to the use of Wikipedia being investigated (see above), it is hoped that as community participation widens, the Project will be able to further develop into a repository which not only points to, but stores standards documents, and asscciated information, as valuable representation information for future preservation. The possibility of further development into a registry for machine-to-machine verification of compliancy could also be considered.

Origins of DCC DIFFUSE Standards Frameworks

The Diffuse Project (Dissemination of InFormal and Formal Useful Specifications and Experiences to Research, Technology Development and Demonstration Communities) was originally funded under the European Commission's Information Society Technologies (IST) 5th Framework Programme and ran from 1 February 2000 until 31 January 2003. Upon completion of the funding for this project, the data created by its partners[15] was retained online as a valuable information resource[16]. Unfortunately no funding was available to maintain the resource and in the rapid moving world of information technology it became outdated. In 2005, the DCC secured permission to re-purpose the content and have redeveloped the concept into the browsable database described above.

The redeveloped DCC DIFFUSE offers the contextualisation of the standards included, by both the specific domains, which find them useful and the lifecycle action that they support. It shows how particular sectors use standards, both domain specific and generic standards together, to achieve their curation aims, and maintain the authority of their digital material. This should enable users a greater understanding of the applicability of standards to their own situation, and enable a more informed choice regarding which to implement. Additionally the harnessing of Wikipedia's volunteers should ensure that the resource does not become so badly outdated if it is not actively maintained for a period.

Conclusions

The DCC has further extended the concept of the original Diffuse Project, to

[15] The partners were: TIEKE (Finnish Information Society Development Centre) - http://www.tieke.fi/in_english/, IC Focus and the SGML Centre (now IS-Thought - http://www.is-thought.co.uk/)

[16] Originally developed in xml, html snapshots from the Diffuse Project are still retained online by IS-Thought http://www.is-thought.co.uk/Diffuse2/home.html

create a newly dynamic resource which enables domain specific standards frameworks, which will support the lifecycle management of authoritative records, to be developed, recorded and managed. Contextualisation, using the DCC Curation Lifecycle Model enables users to identify appropriate standards and the suitable lifecycle stages for their implementation, facilitating detailed curation planning and the realisation of curation activities.

Acknowledgements

I would like to thank both members of the DCC Staff and the various experts who submitted contributions to the development of the DCC Curation Lifecycle Model. Additional thanks go to Chris Blackall for suggestions in respect of lifecycle graphics (and advice on the use of Wikipedia) and Sue Fairhurst of the University of Bath who realised them from my rough sketches.

References

[website] Diffuse Project (2003). *The Diffuse Project Homepage*. Retrieved July 25, 2008 from http://www.is-thought.co.uk/Diffuse2/home.html

[proceedings] Dunning, A., Hollins, P,. Johnston, P., Kelly, B., Phipps, L. and Russell, R. (2005) Standards framework for digital library programmes. *International Conference on Hypermedia and Interactivity in Museums, 2005, Paris, France.* Retrieved July 25, 2008 from http://www.ukoln.ac.uk/web-focus/papers/ichim05/html/

[Internet journal] Gill, T., and Miller, P. (2002, January) Re-inventing the Wheel? Standards, Interoperability and Digital Cultural Content. *D-Lib Magazine 8(1).* Retrieved 25 July 2008 from http://www.dlib.org/dlib/january02/gill/01gill.html#11

[book] Foulonneau, M., and Francis, A. (2007) *Investigative study of standards for Digital Repositories and related services*, Amsterdam, Amsterdam University Press

[Internet journal] Higgins, S. (2008, July). The DCC Curation Lifecycle Model. *International Journal of Digital Curation 3(1).* Retrieved 18 October 2008 from http://www.ijdc.net/ijdc/article/view/69/69

[Internet journal] Hodge, G.M. (2000, Jan). Best practices for archiving: an information life cycle approach. *D-Lib Magazine 6(1).* Retrieved July 24, 2008, from http://www.dlib.org/dlib/january00/01hodge.html

[standards specification] International Organization for Standardization (2001). *ISO 15489-1(2001) Information and documentation -- Records management -- Part 1: General.* Geneva, Switzerland: ISO Publications

[standards specification] International Organization for Standardization (2001). *ISO*

15489-1(2001) Information and documentation -- Records management -- Part 2: Guidelines. Geneva, Switzerland: ISO Publications

[Internet journal] Miller, P., Dawson, D. and Perkins, J. (2001, October) Standing on the shoulders of giants, *Cultivate Interactive, 5(1).* Retrieved July 25, 2008 from http://www.cultivate-int.org/issue5/giants/

[preprint] Pennock, M. (2007). Digital curation: a life-cycle approach to managing and preserving usable digital information. *Library and Archives Journal, Issue 1.* Retrieved July 25, 2008 from http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf

[standards specification] Serco Consulting (2008) *Model Requirements for the management of electronic records update and extension, MoReq2 specification.* Bruxelles- Luxembourg, CECA-CEE-CEEA. Retrieved July 27, 2008 from http://www.moreq2.eu/downloads.htm