Action for oais.review.info:103

This action combines two tasks / comments in the referenced Item:

- Comment 2: **TerryLongstreth** **2017-06-13 16:56:44 CEST** `I will review document for consistent use of term 'Data Model' and report back to DAI.`

- Comment 3: **TerryLongstreth** **2017-06-20 17:50:23 CEST** `Review Data Submission Session description and consistency with Submission Agreement description`

Term 'Data Model'

There are 13 occurrences of the term 'data model' spread throughout the 2012 edition of OAIS. The table below shows the page location and wording/context for each occurrence. I haven't had time to do more than a quick, subjective characterization of the individual sections ('Interpretation') but the list itself clearly illustrates that 'Data Model' is an ambiguous term that doesn't have a simple or consistent meaning in the standard. In some cases, the US government contracting term 'data item description' (DID) could probably have been used in place of data model. There's an implication that Data Models are graph oriented drawings of relationships between data elements or aggregates, but many of the cases below should clearly include either procedural descriptions of functions associated with these relationship graphs, or even non-procedural, predicate-based formulae for covering the expected transformations governed by the 'data models'.
/s/ Terry Longstreth

Interpretation	Page Ref	Wording
First sentence implies Model Views are the larger abstraction. Conceivably, could include process model views. Note term 'data model views' is used in second sentence, not just 'data model'. Implication is that other views of information (perhaps, schemata or ontologies) are anticipated or expected.	**1.6.2 ORGANIZATION BY SECTION** p.1-6	Section 4 provides model views needed for a detailed understanding of an OAIS Archive. It breaks down the OAIS into a number of functional areas and it identifies some high-level services at the interfaces. **It also provides detailed data model views of information using Unified Modeling Language (UML) diagrams.**
Seems data model is synonymous with 'format and contents specification'	**1.7.2 TERMINOLOGY** p.1-10	**Data Dissemination Session**: A delivery of media or a single telecommunications session that provides Data to a Consumer. **The Data Dissemination Session format/contents is based on a data model negotiated between the OAIS and the**

Interpretation	Page Ref	Wording
for the query		**Consumer in the request agreement. This data model identifies the logical constructs used by the OAIS and how they are represented on each media delivery or in the telecommunication session.**
Data model identifies 'logical constructs'. As distinguished from physical, imaginary, natural, or … constructs?	**1.7.2 TERMINOLOGY** p.1-11	**Data Submission Session**: A delivery of media or a single telecommunications session that provides Data to an OAIS. **The Data Submission Session format/contents is based on a data model negotiated between the OAIS and the Producer in the Submission Agreement. This data model identifies the logical constructs used by the Producer and how they are represented on each media delivery or in the telecommunication session.**
Here, the data model is also contrasted with 'other arrangements' needed for data submission session.	**1.7.2 TERMINOLOGY** p.1-15	**Submission Agreement: The agreement reached between an OAIS and the Producer that specifies a data model, and any other arrangements needed, for the Data Submission Session. This data model identifies format/contents and the logical constructs used by the Producer and how they are represented on each media delivery or in a telecommunication session.**
Here, the Data Submission Session appears to be governed by SIP contents and an externally defined 'Data Model' and that data model includes the SIP Schemata as well as definitions of PDI, Repinfo, and Content data objects. It's unclear if the data model is a formally defined schema, or a natural language description, and if it's (fully or partially) machine processable. In the example, it's unclear if the data model governing the Submission Session	**2.3.2 PRODUCER INTERACTION** pp.2-9(-10)	Within the Submission Agreement, one or more Data Submission Sessions are specified. **There may be significant time gaps between the Data Submission Sessions. A Data Submission Session will contain one or more SIPs delivered as a set of media or via a single telecommunications session. The Data Submission Session content is based on a data model negotiated between the OAIS and the Producer in the Submission Agreement. This data model identifies the logical components of the SIP (e.g., the Content Data Objects, Representation Information, PDI, Packaging Information, and Descriptive Information) that are to be provided and how (and whether) they are represented in each Data Submission Session. All data deliveries within a Submission Agreement are recognized as belonging to that Submission Agreement and will generally have a consistent data model, which is specified in the Submission Agreement. For example, a Data**

Interpretation	Page Ref	Wording
(and described in the Submission agreement) contains and encompasses all of the physical specifications for the media themselves, or even if media are an acknowledged part of the Submission Agreement, and how if the Data Model doesn't prescribe specifics of media handling, would that info be communicated.		**Submission Session may consist of a set of Content Information corresponding to a set of observations, which are carried by a set of files on a CD-ROM. The Preservation Description Information is in this example split between two other files. All of these files need Representation Information which must be provided in some way. The CD-ROM and its directory/file structure are the Packaging Information, which provides encapsulation and identification of the Content Information and PDI in the Data Submission Session. The Submission Agreement indicates how the Representation Information for each file is to be provided, how the CD-ROM is to be recognized, how the Packaging Information will be used to identify and encapsulate the SIP Content Information and PDI, and how frequently Data Submission Sessions (e.g., one per month for two years) will occur. It also gives other needed information such as access restrictions to the data and enforcement requirements.**
First mention of a "physical" data model, with no guidance as to how it differs from other uses of term 'data model' (or data model view).	**4.3.3 DATA TRANSFORMATIONS IN THE ARCHIVAL STORAGE AND DATA MANAGEMENT FUNCTIONAL AREAS** p.4-53	The Archival Storage Functional Entity takes the AIPs produced by the Ingest process and merges them into the permanent Archive holdings. The Data Management Functional Entity takes the Package Descriptions produced by Ingest and augments the existing Collection Descriptions to include their contents. The logical model of the ingested data should already be mapped into the logical model of the Archive's holdings. **Thus the major transformation that occurs in this step is the mapping of the acquisition session from the ingest physical data model.** This will tend to be on temporary storage, to the permanent storage of the OAIS, which could range from a Database Management Systems (DBMS) to a Hierarchical File Management Systems (HFMS), or any mixture of the above.
Here, the discussion strikes out in another direction, by contrasting the (partial?) data model describing only Content Information, with a new 'more varied' data model. Why a	**5.1.3.4 Transformation** p.5-8	**The following scenario identifies a Non-Reversible Transformation that can occur when Content Information is migrated to a new format that can express a more varied data model than the original format.** All the Content Information bits for an AIP are contained within the body of three files on a CD-ROM. The

Interpretation	Page Ref	Wording
consequence of a more varied model could result in a loss of data is not explained. An incompatible or sparse output mapping *could result* in data loss, but the issue cannot be described as dependent upon a single data description ('model'). In the last sentence, the issue is further complicated by the example of a non-reversible transformation being defined by the absence of an algorithm, with no explanation of how this relates to data models.		Packaging Information includes the bits used to implement the file and directory structure that provides access to these three files. The contents of the three files are transferred to a new CD-ROM and in the process the third file is altered because there are no longer readily available tools to make effective use of the third file's content in its current form. **The new format, which is in common use, employs a different data model from that of the original format and there are many ways in which the information may be mapped into the new format.** This mapping must be carefully done to ensure there is no significant information loss to the Designated Community. For example, for scientific data a Transformational Information Property could be the values of identified data elements to a specified precision; if the Content Information was a document then the page layout might be a Transformational Information Property. This mapping from the previous format to the new format must be included in the PDI, and of course the Representation Information describing the new format will replace that which was describing the previous format. The result is a new AIP Version. This is a Transformation type of Migration that is also a Non-Reversible Transformation when there is no algorithm that will reproduce the original file from the new file.

Consistency between Submission Agreement and Data Submission Session descriptions

These two terms are shown in the table above, as **1.7.2 TERMINOLOGY** p.1-15 and **1.7.2 TERMINOLOGY** p.1-11, respectively.

It's not clear if a Data Submission Agreement is intended to cover multiple sessions or if a new Agreement is needed for each Session. Other text (if any) referring to the functional intentions of these agreements should be reviewed, and if the issue isn't clarified there, I'd recommend changing the wording of the Submission Agreement to indicate that it will apply to all identified Submission Sessions.