Data Standardisation for Clix Project

Abstract
Clix project at the core is the design and implementation of connected learning spaces with technology and teachers as facilitators. Teams involved - Domain, Implementation, Research and Technology - are working closely in a distributed setting to sucessfully deliver the project. Data is generated and consumed by all the teams and takes special meaning in their respective contexts. There is constant exchange of data among the teams, be it for analysis, reporting or documentation. Unfortunately, as they are generated in different contexts/teams, smooth exchange has become very difficult, slowing down different processes in the project. Also at a broader level, the data generated from this unique intervention needs to be accessible to researchers and educationists across the world. Apart from these, there is stringent requirement from funding agencies and journal publishing houses for the clear documentation of data generated as part of a project proposal. In this document we try to see one possible approach to solve this issue with an example implementation of data standardisation.

Keywords
metadata — data documentation — data repository

contact:Data Engineer, Clix Project, TISS, Mumbai

Contents

Introduction

The core element of good data documentation is content metadata. In this document, we start with a very brief overview of some of the popular metadata standards. Then we will dive into data landscape of Clix project. This is followed by implementation details of DDI (Data Documentation Initiative standard) for a sample data. Finally, we discuss some of the implementation aspects specific to our project.

1. Metadata Standards

Metadata is basically a descriptor of data or any digital object that help us understand what it is, where to find it, how to access it[1]. Different metadata standards emphasize different aspects of data like discoverability, accessibility etc depending on their purpose. Landscape of metadata standards is complex with different standards developed depending on their perspective from the domain they are interested in, motivation for development of standards (filling the gap or combining different existing standards) and also based on the stakeholders involved. Recently, there have been efforts to define standards to make data reusable not just by Human peers but also help machines automatically[2] find and use data.

In this section, we briefly mention few of the popular metadata standards, just to give an idea of their variety. **For any detailed understanding of standards and data documentation in general please refer references mentioned at the end of this document.**

DDI Data Documentation Initiative[3] is a free international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences. DDI is a free standard that can document and manage different stages in the research data lifecycle, such as conceptualization, collection, processing, distribution, discovery and archiving. Documenting data with DDI facilitates understanding, interpretation and use – by people, software systems and computer networks.

As far as implementation[4] is concerned, it is as simple as editing XML files following a standard schema(called codebooks). There are many tools available for creating DDI documents to accompanying data, specifically for windows operating system. We may have to edit raw XML files using editors in linux systems.

Dublin Core Metadata Initiative[5] They design, manage and curate different metadata standards across different disciplines. Popular standards maintained by them are Datacite metadata schema, Darwin core germplasm etc.

Resource Description Framework Schema[6] It provides vocabulary for concpetual description of resources on web. It

is developed by W3C and is widely used for resources on web and actively maintained.

MARC standards[7] The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form.

ODPI Egeria[8] This is a first open source enterprise level metadata standards specification, developed and delivered under Apache 2.0 open source license. The projects objective is to develop open source solutions for easy access and governance of data across different verticals of an enterprise in Big Data ecosystem. It is specifically suited for enterprises in the big data business.

2. Data Landscape of Clix Project

This section gives a broad overview of different sources and modalities of data generated as part of Clix intervention. Data generated from project can be broadly divided into survey data and platform data. Each of these are explained in some detail below:

2.1 Survey Data

This corresponds to survey and interviews conducted by clix team at a school level during different stages of the project to understand various aspects of the intervention - adoption and learning outcome. They comprise of carefully designed questions for teachers, principals, filed implementors and classroom observors.

Baseline-Endline study This corresponds to field survey designed for principals, teachers and students to understand changes in their attitudes towards intervention. They also include Innovation Diffusion Process Documentation (IDPD) interviews involving field implementation groups and classroom observations by researchers.

Pre-Clix and Post-Clix survey This corresponds to student survey before and after exploring the Clix modules and is available from the platform, as and when students use the module.

Implementation Monitoring Tool This data corresponds to periodic input from the field implementation teams on adoption of the project at a school level.

2.2 Platform Data

This comprises of all the data collected from Clix software (at machine level in any given school). It has many component data logs each corresponding to different aspect of the platform. Below is a brief description of the same[9],[10].

Course Player Data This data is generated as the user (student or non-student) navigates through the platform to use course's module-unit-lesson components, by periodically logging his/her interaction. The data is at the unit level and is in the form of csv.

Interactive Tools Data Data here corresponds to student's access of third party tools(Apps) in the platform. These are json log files generated from tools section of the platform, whenever a user logs into clix platform and navigates to tools section these log files are generated. This data is also very granular in the sense that detailed interaction of student with the tool is captured in json files.

Benchmark Data This corresponds to data captured at the activity level of course modules. It directly accesses node (in the context of gstudio software architecture) level information. The data is in the form of time-stamped csv files (activity time stamp files) and consist of detailed logs of interaction of student with the activity pages of the course module.

MIT Tools Data These consist of open student assessments and also interaction with other tools (run-kitty-run and open story tool) developed outside gstudio framework. These are logged as json files in Qbank part of Mongodb (which is a component of Clix software in every machine).

User Artefacts Data These are user generated artefacts through clix platform. They can be files uploaded, ratings given to other resources, making notes, giving feedbacks etc. Cookies are used to get more of this information in gstudio modules. As of now there is no direct way of accessing this data.

3. DDI Implementation Example

This section describes important components in the implementation of DDI standards. At the very basic level, DDI implementation can be thought of as starting with a text document of description of study/data (metadata) and then tagging every information in that text document with a predefined tag (so that information is given a context now and machines can also make sense of them). DDI provides guidelines for metadata specification and requires to implement these standards in XML format. The specifications have two variants - DDI codebook and DDI life cycle. Codebook is just a lighter version of the Life Cycle, intended to document simple survey data. DDI Life Cycle includes entire life cycle of data management from conceptualization to publishing. In the rest of the section, we will go through implementation details of Platform Adoption Study data (an example dataset prepared for illustration puposes) using Codebook2.0[11] specifications of DDI.

Platform Adoption Study Data - Codebook The codebook specification has four components - Documnet description, Study description, Data files description and Variable description. We start with a text document collating all the information required for these components:

Document Description Here the DDI document itself is described and possible entries here are Citation, Document Source

Study Description This is description of the study. Citation, Study Scope, Methodology and Processing, Data Access, Other Study Description Materials are included here

File Description Data files being documented are described in this part of the xml file. Fields here include Data file content information

Variable Description All the variables involved in the data set are documented here. Variable Groups and Individual variables are its component fields

Finally, we encode this information in xml format using available xml editors. With this, we have a fully functional DDI file to accompany our adoption study data. This can be made browser friendly by adding xsl stylesheet. This was an example of cross-section data of student's usage pattern of tools section of the platform, same can be applied to survey data also.

Note please find ddi-xml file, xsl stylesheet and actual data set at this **folder**. Just open xml file in Firefox with all the files in same folder.

4. Implementation Aspects Specific to Clix

In this section, we will describe possible workflows involved to standardise data using DDI specifications in the Clix project setting. Before we begin to think about workflow, lets reflect a bit on objective of standardisation.

Objective Whenever a team generates a data set, it has to be properly documented (as per DDI standards) for it to be used by other teams and may be researchers outside the institution. And it should become part of Clix data repository, which can be easily accessed by everyone.

Now standardization requires three core ingredients - *knowledge of the data (to define the context), knowledge of the standard specifications (to decide on which information to document) and finally techincal ability to put the final documentation in XML format*. Please note that, here we have to distinguish between raw data (which is just data generated as digital exhaust or part of project implemtation) and data generated/created by teams (which has some purpose). Standardisation applies to the former case and raw data is just stored and archived following standard data practices.

Given this, we may start with defining schema (which defines columns to be included in a dataset created from raw data) for all the datasets of interest. Then systematically create DDI documents for these datasets over the span of a month or two and finally push them to a repository (an easily accessible database or file location). Creating a DDI document has some initial costs (in terms of time-spent) involved, but this can be ammortized over subsequent efforts through automation as xml files are machine accessible.

5. Summary

To summarize, we tried to recognize the issue of standardisation of data in the Clix project context. Explored possible alternatives to address this and saw an implementation of one of the appproaches (DDI) with an example data. Finally, noted key components for successful implementation of the same at the project level.

References

[1] Dr. Susanna-Assunta Sansone. BD2K Guide to the Fundamentals of Data Science. **Metadata Standards**, 2016.

[2] M. D. et al. Wilkinson. The fair guiding principles for scientific data management and stewardship. *Scientific Data, Nature*, Sci. Data 3:160018, 2016.

[3] Data Documentation Initiative **Website**.

[4] ICPSR. Guide to social science data preperation and archiving - best practice through data life cycle. *Institute for Social Research, University of Michigan*, 5th edition, 2012.

[5] Dublin Core Metadata Initiative **Website**.

[6] Resource Description Framework Schema **website**.

[7] MARC framework **website**.

[8] ODPI Egeria **website**.

[9] Clix Data Walkthrough **document**.

[10] Clix Data Almanac **document**.

[11] DDI Codebook Examples **Website**.