The Standardization Survival Kit: for a Wider Use of Metadata Standards within Arts and Humanities

Charles Riondet, Laurent Romary

THE STANDARDIZATION SURVIVAL KIT: FOR A WIDER USE OF METADATA STANDARDS WITHIN ARTS AND HUMANITIES

Charles RIONDET, Laurent ROMARY

Introduction

In this paper, we would like to set out an innovative way of documenting best practices and protocols related to research processes in Arts and Humanities: the Standardization Survival Kit (SSK), an open platform for hosting resources related to standards, curated in research scenarios. One of the main goals of this platform is to offer guidelines for metadata creation and management. We illustrate this tool with a scenario that offers a workflow for managing heterogeneity of archival standards, in particular the *Encoded Archival Description* (EAD), based on a customisable framework, TEI-ODD (*Text Encoding Initiative-One Document Does it all*).

European research infrastructures such as DARIAH, PARTHENOS or EHRI play a key role to set out policies and give the framework for creating interoperable and sustainable resources. Availability of the content is partially an infrastructural task, but standards are very often the pivot of such policies or endeavours. The work presented here has been carried out for the PARTHENOS project[1], which aims at harmonizing as much as possible the practices and policies of researchers in different Humanities disciplines, by giving access to an integrated environment of research services and tools. It is also related to another European project: EHRI[2], and more specifically its main product, the EHRI portal, giving access to more than 19,000 archival descriptions related to the Holocaust.

In such contexts, proper data modelling and corresponding standards play a crucial role, to make digital content more sustainable and reusable. In a research process (individual or collective), there is always an initial phase during which researchers should be made aware of some core and domain oriented standards, in order to prevent the specification of *ad hoc* local formats overlapping with existing sustainable solutions available in the Digital Humanities landscape. As there is no formal obligation to follow a standard when doing research, except when one

[1] http://www.parthenos-project.eu.

[2] L. Romary and C. Riondet, "EAD-ODD: A solution for project-specific EAD schemes", in *Archival Science*, 2018, https://dx.doi.org/10.1007/s10502-018-9290-y. hal-01737568v2.

actually wants to produce comparable and verifiable results, it is one of the endeavours of an infrastructure in the Arts and Humanities to recommend best practices for the scholarly communities regarding the adoption and the implementation of specific standards. It is essential to provide potential users with an awareness of the appropriate standards and the advantages to be gained by adopting them[3], but it is even more crucial to present the cognitive tools to help them identify the optimal use of standards through the selection and possibly customisation of a reference portfolio.

To help researchers and research teams addressing these issues, PARTHENOS has identified the notion of the *Standardization Survival Kit* or SSK[4]. A secondary goal is to foster innovative, cross-disciplinary research paths able to bridge the gaps existing between the different *episteme* that compose the broad landscape of Humanities and Cultural Heritage studies.

The Standardization Survival Kit

The SSK aims at giving answers about four types of activities related to the deployment and use of standards in the Humanities and Cultural Heritage fields:

1. It documents existing standards, by providing reference material for scholars who want to find out more about their role and content. This relates to the specific provision of bibliographic sources, available documentation, specific targeted introductions, as well as providing prototypical examples which can serve as models for similar work.
2. It supports the actual adoption of standards by identifying how they relate to research scenarios and gathering the essential materials for controlling their deployment (e.g. schemas).
3. The SSK is also a communication tool with research communities so that they can be made aware of both the need to apply standards in their digital scholarly practices but also be informed of the essential standards for their own fields.
4. Last, it is a training tool for researchers, by giving them access to complete frameworks so that they may acquire knowledge and know-how on standardized methodologies.

To cover these four aspects, it appears that the best way to proceed is working on the basis of contextualized use cases, in order to give researchers access to standards in a meaningful way. That is why the core of the SSK is the notion of research scenario. A scenario provides contextual information and relevant examples on how standards can be applied in a given research project. The SSK

[3] L. Romary, "Stabilizing knowledge through standards - A perspective for the humanities. Karl Grandin. Going Digital: Evolutionary and Revolutionary Aspects of Digitization", in *Science History Publications*, 2011. https://hal.inria.fr/inria-00531019.

[4] http://ssk.huma-num.fr.

intends to host scenarios that cover most domains of the Humanities, from Literature to Heritage science, including History, Social sciences, Linguistics, etc. They are derived from real life researcher-oriented use cases and written by domain experts[5]. A scenario is a set of research steps involving specific tasks. They can be seen as a living memory of what should be the best research practices in a given community, made accessible and reusable for other researchers wishing to carry out a similar project but unfamiliar with the recommended tools, formats, methods to use, etc. For that reason, the SSK can be considered as a complete methodological framework showing concrete use of standards, rather than simply a catalogue of resources.

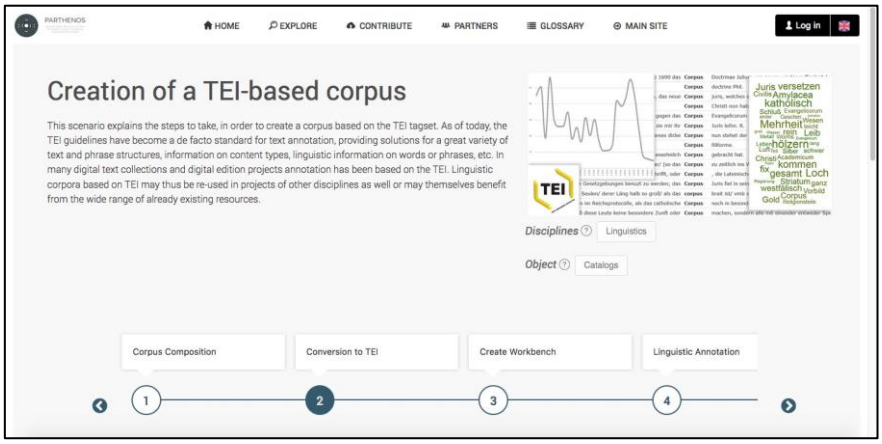

Illustration 1: The SSK: example of a scenario

Each scenario within the SSK works like a high-level research guide for scholars to be followed as a complete process to solve a given problem with the most standardized means. For each step, the action and methodology to follow are described in natural language, and exemplified with appropriate resources. These resources can be generic (the primary documentation and tools) or project-specific (pointing to concrete use cases in which a similar task was accomplished). By resources, we mean first and foremost a state-of-the-art bibliography, which includes all the documentation needed to carry out a given task. The bibliographical references are up-to-date and gathered within a Zotero library[6], which was specially created for this project. The SSK also offers more technical resources, such as stylesheets, code samples, software or services, and training materials like tutorials.

Moreover, we went one step further to be consistent with our principles down to the details. We used a standard, the TEI, as the actual SSK underlying data model

[5] L. Romary, E. Degl'Innocenti, C. Riondet, K. Illmayer, A. Joffres et al., *Standardization survival kit (Draft)*, 2016, https://hal.inria.fr/hal-01513531.
[6] https://www.zotero.org/groups/427927/ssk-parthenos.

for describing all parametrizable content, adopting and customizing the <*event*> element[7]. Following this same idea, all the data is publicly available in a GitHub repository[8]. Every scenarios and steps are encoded in TEI documents, linked together with referencing mechanisms. This mechanism allows for reuse and customization of all the scenarios and steps: The data model allows scenario creators, or any other user of the SSK, to modify the structure of their research scenarios on the fly, by creating, removing or reordering steps. As steps are considered as autonomous objects in the architecture, they can be used in several scenarios.

Metadata standards and the SSK: the example of EAD

Giving access to meaningful and interoperable metadata is one of the fundamental components of open and sustainable research and research data management. The SSK therefore reflects this by presenting scenarios dedicated to metadata, from best practices for creation of generic metadata to the advanced use of bibliographic metadata for research purposes[9]. But in all research scenarios, one or several steps tackle the topic of describing the data involved, produced, derived, *etc.* The metadata standards mentioned in the scenarios are either horizontal standards, i.e. usable in many different contexts, like Dublin-core, CIDOC-CRM, METS, or vertical standards, focusing on very specific data and very specific communities. We can cite bibliographical or archival standards, like EAD, EAC-CPF or EAG.

In all these cases, the metadata standards are promoted as instruments for performing efficient and sustainable research processes. Within the SSK, they are cited in context, with examples taken from existing projects. In some cases, the scenario itself presents how metadata are produced or manipulated in a given project. This is the case for the scenario "Project-centred EAD customization"[10], which presents the methodology and the resources followed by the European Holocaust Research Infrastructure (EHRI), in order to create a specific profile for EAD, used for data integration, enrichment and sharing life cycles[11]. EHRI main product is a federated portal gathering dispersed sources about the Holocaust, hosted by more that 1900 institutions with different histories of custody, cataloguing practices or digitization level[12]. The final goal is to ingest all the relevant archival descriptions in a single environment, with EAD2002 as the pivot format, but with EHRI specific description rules. A pending question, related to

[7] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-event.html.
[8] https://github.com/ParthenosWP4/SSK.
[9] http://ssk.huma-num.fr/#/scenarios/SSK_sc_DisseminationFieldSurveys/1, http://ssk.huma-num.fr/#/scenarios/SSK_sc_trackingTheDisseminationV2/1, http://ssk.huma-num.fr/#/scenarios/SSK_sc_creatingMetadata/1.
[10] http://ssk.huma-num.fr/#/scenarios/SSK_sc_schemaCustomization/1
[11] C. Riondet, L. Romary, A. van Nispen, K. J. Rodriguez and M. Bryant, "Report on Standards", 2017, https://hal.inria.fr/hal-01503235.
[12] https://portal.ehri-project.eu/.

EAD permissive nature is how to preserve content and meaning when exchanging or reusing archival content? This scenario proposes a system that narrows EAD permissiveness and allows for quality checks and content-oriented rules, without modifying the original EAD 2002 schema. Below are explained the main steps, as they can be found in the SSK.

Step 1: Express an XML schema with ODD

It is based on the TEI long lasting experience of customization and specification, in particular its subset called One document does it all, or ODD[13], a very powerful system, with which it is possible to model specific subsets, extensions or profiles of the described format. In other words, it allows us to refine easily the behaviour of elements and attributes, for any XML format, can contain all the human readable documentation and can be processed to generate various resources: a validation schema (many formats) and associated documentation (many formats).

In ODD, documentation (descriptive elements like <gloss> and <desc>) and technical specification (available attributes, authorized children, *etc.*) are brought together, as showed by the following code fragment:

```
<elementSpec ident="unittitle" module="EAD">
        <gloss>Title of the Unit</gloss>
        <desc>The name, either formal or supplied, of the described
        materials...</desc>
        <classes>
                <memberOf key="att.EADGlobal"/>
        </classes>
        <content autoPrefix="true">
                <alternate minOccurs="0" maxOccurs="unbounded">
                        <textNode/>
                        <elementRef key="unitdate"/>
                        <elementRef key="num"/>
                        <elementRef key="date"/>
                </alternate>
        </content>
        <exemplum/>
        <remarks>
                <p>Do not confuse <gi>unittitle</gi> with Title
                <gi>title</gi>,...</p>
```

[13] S. Rahtz and L. Burnard, "Reviewing the TEI ODD System", in *Proceedings of the 2013 ACM Symposium on Document Engineering*, p. 193–196. DocEng '13. New York, USA: ACM, 2013. https://doi.org/10.1145/2494266.2494321.

We created an ODD file that covers EAD entirely, maintained by the project PARTHENOS, with the possibility to contribute and reuse[14].

Step 2: Express projects requirements in machine-readable format
Furthermore, using ODD allow one to derive a specific EAD profile by adding very precise content oriented rules. Each new EAD profile requires a new ODD, that inherits from the master source but allows for the possibility to modify the elements that have a different behaviour. The following illustration shows schematically how the derivation mechanism works and how an element can be modified with a derived, or chained, ODD.

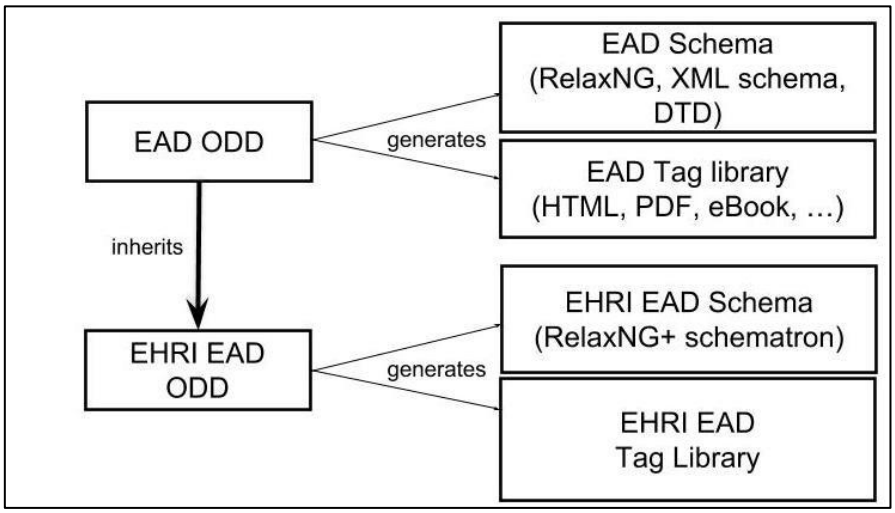

Illustration 2: ODD Derivation mechanism

It is also possible to keep the core EAD schema as it is, and add more specific rules in another validation language, for instance Schematron, an ISO standard used to make assertions or report the presence and absence of XML patterns[15]. EHRI used Schematron rules in three identified cases:

- Emphasize EAD validation errors: Content normalisation (dates, codes, …)
- Align the descriptions with project constraints

[14] http://github.com/ParthenosWP4/standardsLibrary/blob/master/archivalDescription/EAD/ODD/EADSpec.xml.
15 ISO/IEC 19757-3:2016. Information technology -- Document Schema Definition Languages (DSDL) -- Part 3: Rule-based validation – Schematron, http://standards.iso.org/ittf/PubliclyAvailableStandards/c055982_ISO_IEC_19757-3_2016.zip.

- Highlight some description elements that could be improved: Not errors, but pieces of advice. In particular for content related elements (existence of copies of the material, bibliographical references, …)

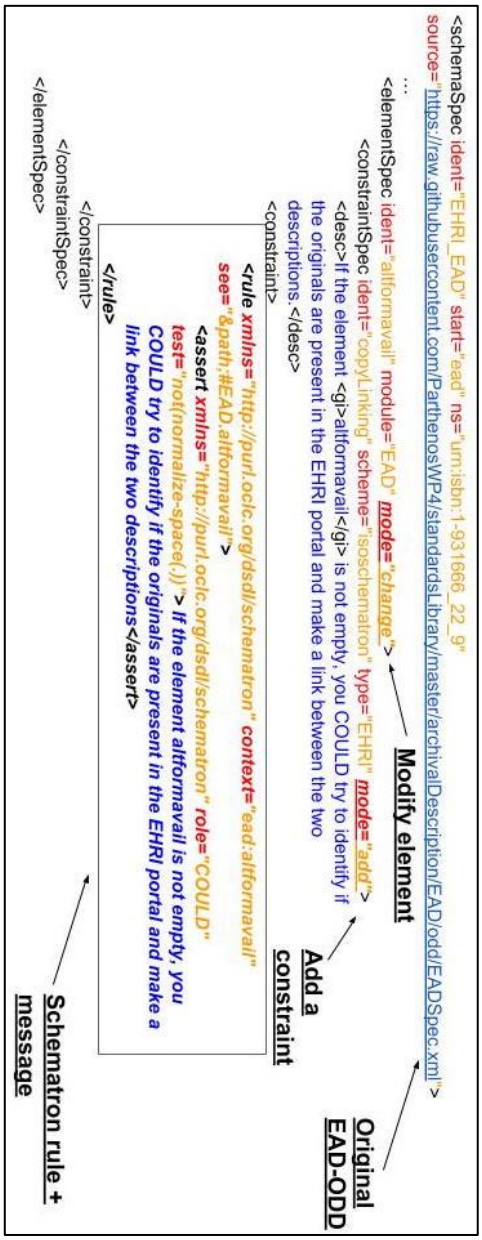

Illustration 3: Modify an element with ODD

Step 3: Create associated documentation
Projects like EHRI follow the best practice of creating comprehensive documentation. We took advantage of this by integrating the human readable documentation in the validation process, and deepen the relationship between validation and documentation.

The documentation of constraints, written by EHRI metadata experts, has been integrated in the ODD file. A full description of the expected content (i.e. HTML "tag library") is generated from the ODD file and in the validation process, the documentation is served to the user in context, whether it is EAD technical documentation, EHRI technical documentation or EHRI qualitative documentation. This makes the validation workflow both technically accurate and easy to understand by the metadata provider.

The fourth and fifth steps of the scenario are the usual tests and publishing phases that occur in all projects of this kind. In EHRI's case, they are integrated in a complete validation and publication process, as shown by the following illustration.

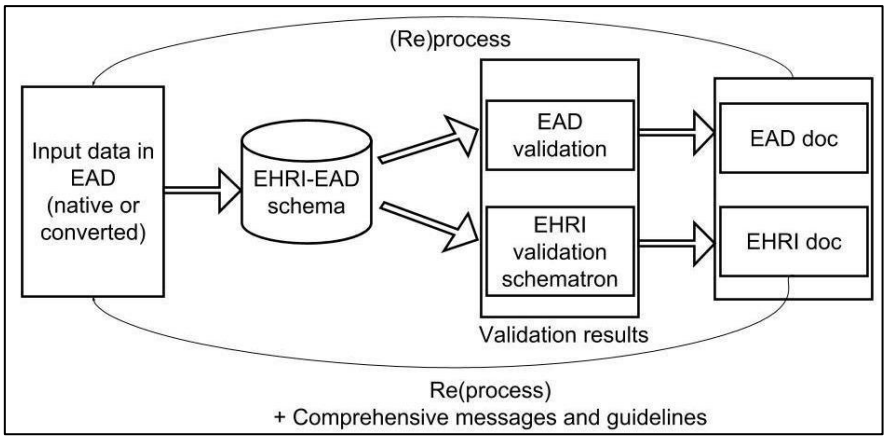

Illustration 4: EHRI's EAD validation process

Conclusion
This paper intends to connect several initiatives that have the use of standardized metadata for commonality. First, we devised a technical solution implementing two standards for managing metadata heterogeneity. Second, we showed how this solution is itself disseminated in a standardized way, exemplifying the use of a new tool for information exchange on standards and data interoperability in Arts and Humanities: the Standardization Survival Kit (SSK). Last, we give a concrete example of collaboration between European Research Infrastructures (PARTHENOS and EHRI, and also DARIAH in the background), that not only try to communicate, but closely connect in order to 1) address their own research issues and 2) share expertise and knowledge.