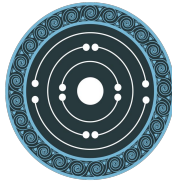


The Standardization Survival Kit (SSK)

Bringing best practices to research communities in the Humanities



PARTHENOS



● What are standards?

3 keywords

- They express a **consensus**
- They are published and easily **accessible**
- They are **maintained**

Data formats

- **XML** :
<https://www.w3.org/TR/xml/>
- TIFF : ISO 12639:1998

Protocols

- **ISO 11554:2017** : Test methods for laser beam power, energy and temporal characteristics

● What are standards?

Present at each step of the research process (examples)

- **Production**: ALTO XML, standard for recording layout and logical structure of OCR'd text;
- **Processing**: Linguistic annotation encoded in XML TEI (Text Encoding Initiative);
- **Archiving**: OAIS (Open Archival Information System, ISO 14721:2012)
=> conceptual model dedicated to the management, archiving and long-term preservation of digital documents.

● Parthenos & the SSK

Initial goal of the *Standardization Survival Kit* in
PARTHENOS :

**Support and provide expertise to researchers in their use
of standards**

Which communities?

- Humanities
- Social Sciences
- Heritage Science

Parthenos & the SSK

Positive context of an European project:

- **Diversity:** many experts from different disciplines
- **Synergy:** willingness to cooperate

⇒ Opportunity to build use cases, inspired from real life.

Parthenos & the SSK

Development team

Inria ALMAAnaCH

- Laurent Romary (DR, WP leader, supervisor)
- Marie Puren, Charles Riondet (project management, data model)
- Dorian Seillier (UI/UX Design)
- Lionel Tadjou, Damien Biabiany (development web)

Iterations, beta tests

PARTHENOS WP4 on Standardization:

- Klaus Illmayer (OEAW),
- Karolien Verbrugge (NIOD),
- Roberta Giacomi (SISMEL),
- Panos Siozos (FORTH),
- and many more

Concept Evolution

1. Support and provide expertise to researchers in their use of standards;
2. Give context to standards;
3. Link them to a concrete research activity ([TaDirah – Taxonomy of Digital Research Activities in the Humanities](#));
4. Link activities between them -> Describe a research process built on the use of standards.

Concept Evolution

Not every research step is related to a standard:

- Ethics and legal issues;
- Evaluation and comparison of results;
- ...

But the notion of **Best practices** is always relevant

● Concept Evolution

When standards are protocols :

- ◉ Normative texts are very formal documents, difficult to read;
- ◉ A protocol = A suite of tasks using tools and techniques.

● Concept Evolution

1. Support and provide expertise to researchers in their use of standards;
2. Give context to standards;
3. Link them to a concrete research activity ([TaDirah – Taxonomy of Digital Research Activities in the Humanities](#));
4. Link activities between them -> Describe a research process built on the use of standards.
5. **A platform for:**
 - a. Documenting research best practices in digital environment
 - b. "Human readable" expression of protocol standards

In PARTHENOS and beyond

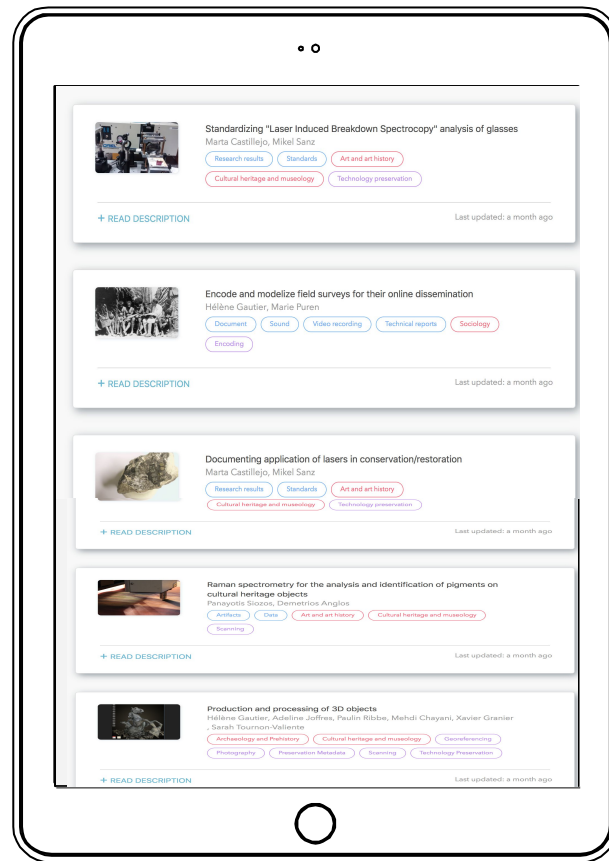
Propose a pleasant, sustainable and adaptable service

- UI/UX designer
- "agile" method: brainstorming, user/usability tests
- "soft" modeling of data
- Meet the potential user communities.



The scenarios

Providing **contextual information** and **relevant examples** on how standards can be applied in a given research project.



Three layers



Scenario

A complete and generic research use case composed of several steps to be followed.



Step

A unique task to be performed inside a scenario with the help and recommendation of one or several resources.



Resource

A standardized tool / service / document guiding the researcher in her/his tasks completion.



3-level Structure

Scenario description (*techniques, disciplines, objets*)

1. Step description (*activities, standards*)
 - a. resource
 - b. resource
2. Step description (*activities, standards*)
 - a. resource
 - b. resource
3. Step description (*activities, standards*)
 - a. etc.

Resources about standards

Standards documentation
(ISO, TEI)

Official publications and
reports (D4Science, HAL,
Zenodo)

Documentation



Reference libraries
organised by domains and
standards

Maintenance with Zotero

Bibliography



Code snippets (GitHub)

Tools & services
(D4Science)

Technical
resources



Wikis

Blog posts
(Hypotheses.org)

Discussion lists

User
communities



[illegible]

Musicality And Performing Arts



Sound

Score

Multiple Score Formats

Create a digital corpus of musical compositions.

Transcription

Editing

Change format into MEI files if necessary.

Conversion





Resources and best practices

musical edition with critical apparatus metadata

General Resources

General resources are made of primary documentation, guidelines and tools.

⑥ Enrichment of MEI files

MEIMessaging

Django/python powered MEI messaging app

Author(s): DuChemin

URL : [SafeValue must use \[property\]=binding: SafeValue must use \[property\]=binding: SafeValue must use \[property\]=binding: https://github.com/DuChemin/MEIMessaging \(see http://q.co/nq/security#xss\) \(see http://q.co/nq/security#xss\) \(see http://q.co/nq/security#xss\)](#)

meiView

JavaScript demo application displaying variant readings of historical musical pieces - zolaemil/meiView

Author(s): zolaemil

URL : [SafeValue must use \[property\]=binding: SafeValue must use \[property\]=binding: SafeValue must use \[property\]=binding: https://github.com/zolaemil/meiView \(see http://q.co/nq/security#xss\) \(see http://q.co/nq/security#xss\) \(see http://q.co/nq/security#xss\)](#)

MEItoVexFlow

JavaScript library to render MEI notation using VexFlow - TEI-Music-SIG/MEItoVexFlow

Author(s): TEI-Music-SIG

URL : [SafeValue must use \[property\]=binding: SafeValue must use \[property\]=binding: SafeValue must use \[property\]=binding: https://github.com/TEI-Music-SIG/MEItoVexFlow \(see http://q.co/nq/security#xss\) \(see http://q.co/nq/security#xss\) \(see http://q.co/nq/security#xss\)](#)



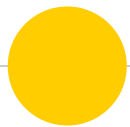
Easy-to-use & collaborative platform

- Consult and follow the guidelines expressed in the scenarios
- Propose new scenarios



Why contributing ?

- Make your research project align with the best practices in your community
- Get peer review and visibility
- Share a project in another form than the usual blog/article (a new way to disseminate your work)



Coherence with our principles

The SSK, a 100% standards web app

Scenarios in TEI

The scenarios are described using the TEI format (Text Encoding Initiative). All the information displayed within the SSK proceed from TEI files.

<http://github.com/ParthenosWP4/SSK/tree/master>

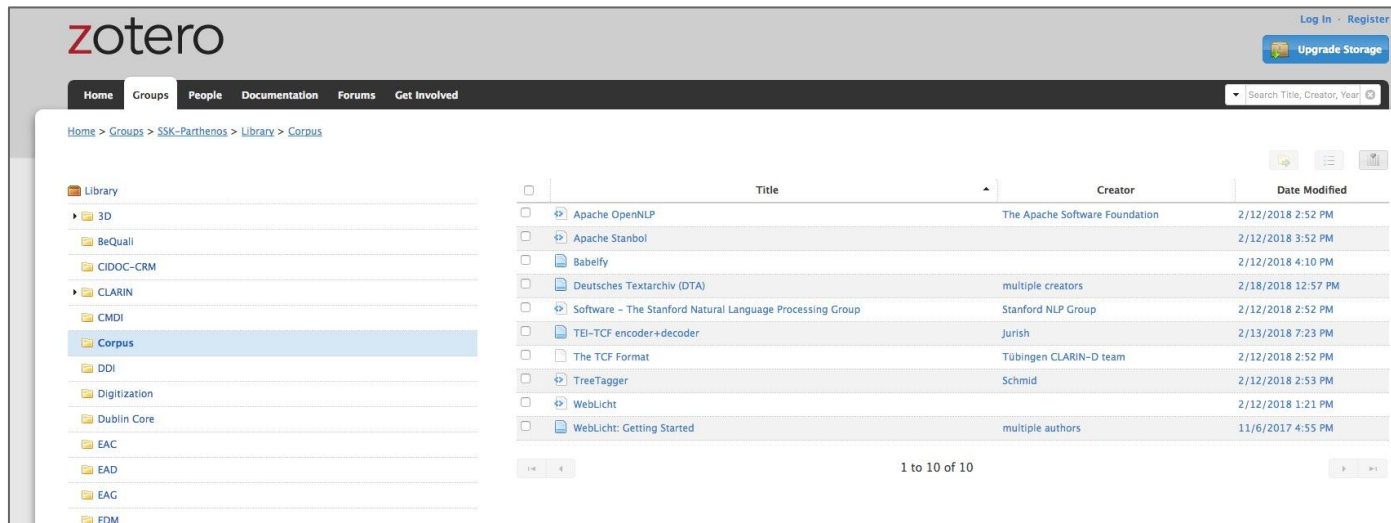
```
<text>
<body>
  <div type="scenario" xml:id="sc_schemaCustomization">
    <head xml:lang="en" type="scenarioTitle">Harmonization of digitized textual resources
      with the DTABf</head>
    <desc type="definition" xml:lang="en">Currently, initiatives for the digitization of
      textual resources and their provision to the interested community are manifold and
      various. Hence, scholars who want to base their research on digitized texts,
```

```
<listEvent>
  <event ref="step_OtICoP_171117" type="researchStep" xml:id="s1"/>
  <event ref="step_CaC_171117" type="researchStep" xml:id="s2"/>
  <event ref="step_SaD_171117" type="researchStep" xml:id="s3"/>
  <event ref="step_A_171117" type="researchStep" xml:id="s4"/>
  <event ref="step_CiSF_171117" type="researchStep" xml:id="s5"/>
  <event ref="step_TtI_171117" type="researchStep" xml:id="s6"/>
```

Managing resources with bibliographic standards

All the **references** are managed by the open source management software **Zotero**, and can be found in the a dedicated library.

<https://www.zotero.org/groups/427927/ssk-parthenos/>

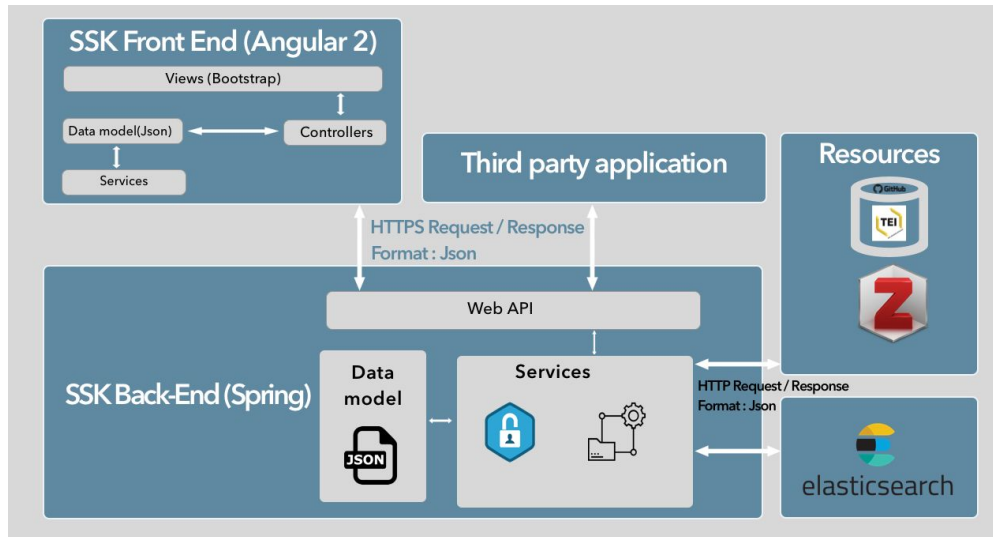


The screenshot displays the Zotero web interface. At the top, the Zotero logo is on the left, and 'Log In' and 'Register' links are on the right. Below the logo is a navigation bar with links: Home, Groups, People, Documentation, Forums, and Get Involved. A search bar is located on the right side of this bar. The main content area shows a breadcrumb trail: Home > Groups > SSK-Parthenos > Library > Corpus. On the left, a sidebar lists various libraries, with 'Corpus' selected and highlighted. The main area displays a table of resources with columns for Title, Creator, and Date Modified. The table contains 10 entries, including 'Apache OpenNLP', 'Apache Stanbol', 'Babelify', 'Deutsches Textarchiv (DTA)', 'Software - The Stanford Natural Language Processing Group', 'TEI-TCF encoder+decoder', 'The TCF Format', 'TreeTagger', 'WebLicht', and 'WebLicht: Getting Started'. At the bottom, a pagination bar shows '1 to 10 of 10'.

Title	Creator	Date Modified
Apache OpenNLP	The Apache Software Foundation	2/12/2018 2:52 PM
Apache Stanbol		2/12/2018 3:52 PM
Babelify		2/12/2018 4:10 PM
Deutsches Textarchiv (DTA)	multiple creators	2/18/2018 12:57 PM
Software - The Stanford Natural Language Processing Group	Stanford NLP Group	2/12/2018 2:52 PM
TEI-TCF encoder+decoder	Jurish	2/13/2018 7:23 PM
The TCF Format	Tübingen CLARIN-D team	2/12/2018 2:52 PM
TreeTagger	Schmid	2/12/2018 2:53 PM
WebLicht		2/12/2018 1:21 PM
WebLicht: Getting Started	multiple authors	11/6/2017 4:55 PM

RESTful architecture

<http://github.com/ParthenosWP4/SSK/tree/dev>



- Flexible, easy to deploy and maintain architecture
- Independent entities communicating via REST services.



SSK data in details

Same model for describing scenarios and steps

<i>Information</i>
Authors
Title
Description
Keywords



SSK data in details

<i>Information</i>	<i>Tei element</i>
Authors	titleStmt/author/
Title	head
Description	desc type="definition"
Keywords	desc type="terms"



SSK data in details

Same model for describing scenarios and steps

But in different contexts

The scenarios description is "above" the steps description

Scenarios in TEI

The scenarios are described using the TEI format (Text Encoding Initiative). All the information displayed within the SSK proceed from TEI files.

A scenario is a list of events (`<tei:listEvent>`), each step in a scenario is an event (`<tei:event>`).

```
<text>
<body>
  <div type="scenario" xml:id="sc_schemaCustomization">
    <head xml:lang="en" type="scenarioTitle">Harmoniza
      with the DTABf</head>
    <desc type="definition" xml:lang="en">Currently, i
      textual resources and their provision to the ir
      various. Hence, scholars who want to base their
```

```
<listEvent>
  <event ref="step_OtICoP_171117" type="researchStep" xml:id="s1">
  <event ref="step_CaC_171117" type="researchStep" xml:id="s2"/>
  <event ref="step_SaD_171117" type="researchStep" xml:id="s3"/>
  <event ref="step_A_171117" type="researchStep" xml:id="s4"/>
  <event ref="step_CiSF_171117" type="researchStep" xml:id="s5"/>
  <event ref="step_TtI_171117" type="researchStep" xml:id="s6"/>
```

Scenarios in TEI

```
<text>
<body>
  <div type="researchScenario">
    <head type="scenarioTitle" xml:lang="en">Creation of a TEI-based corpus</head>
    <desc type="definition" xml:lang="en">This scenario explains the steps to take, in order to
      create a corpus based on the TEI tagset. As of today, the TEI guidelines have become a de
      facto standard for text annotation, providing solutions for a great variety of text and
      phrase structures, information on content types, linguistic information on words or
      phrases, etc. In many digital text collections and digital edition projects annotation has
      been based on the TEI. Linguistic corpora based on TEI may thus be re-used in projects of
      other disciplines as well or may themselves benefit from the wide range of already
      existing resources.</desc>
    <desc type="terms" xml:lang="en">
      <term source="aurehal" type="discipline">Linguistics</term>
      <term key="text" source="Tadirah" type="object"/>
    </desc>
    <figure type="image">
      <head>Illustrative image of the scenario</head>
      <graphic
        url="https://raw.githubusercontent.com/ParthenosWP4/SSK/master/img/corpusAnalysis_compact.png"/>
      <figDesc>Combination of screenshots of different steps in the scenario (image created by
        Susanne Haaf)</figDesc>
    </figure>
    <listEvent>
      <event ref="step_corpusComposition" type="researchStep" xml:id="s1"/>
      <event ref="step_verificationAndCleanup" type="researchStep" xml:id="s2"/>
      <event ref="step_conversionToTEI" type="researchStep" xml:id="s3"/>
```

Steps in TEI

```
<body>
  <listEvent>
    <event type="researchStep">
      <head type="stepTitle" xml:lang="en">Conversion to TEI</head>
      <desc type="definition" xml:lang="en"> For the project at hand, a TEI format has
        to be chosen or created (the latter by usage of the ODD language) which
        suits the markup necessities defined in the corpus composition step. Thus,
        if digitized data from other sources are to be re-used for corpus creation,
        these may very likely be available only in formats that aren't similar to
        the TEI format selected for the corpus creation project at hand. External
        data may either come in completely different formats or at least in
        different TEI dialects. In any case, it will be necessary to convert the
        data from different formats into the TEI output format. Conversion may be
        conducted semi-automatically.</desc>
      <desc xml:lang="en" type="terms">
        <term type="standard" source="standard_list" key="XML"/>
        <term type="standard" source="standard_list" key="TEI"/>
        <term type="activity" source="http://tadirah.dariah.eu/" key="conversion">Conversion</term>
      </desc>
      <linkGrp type="generalResources">
        <ref type="spec" source="zotero" target="ZABRV5VD">
          <term type="standard" source="standard_list" key="TEI"/>
        </ref>
        <ref type="spec" source="zotero" target="ZE34VR34">
          <term type="standard" source="standard_list" key="TEI"/>
        </ref>
        <ref type="spec" source="zotero" target="ZKBT7U3E">
```



Keywords in TEI

<term> element. Doc: https://sisk.readthedocs.io/en/latest/2_sisktei.html#term-element

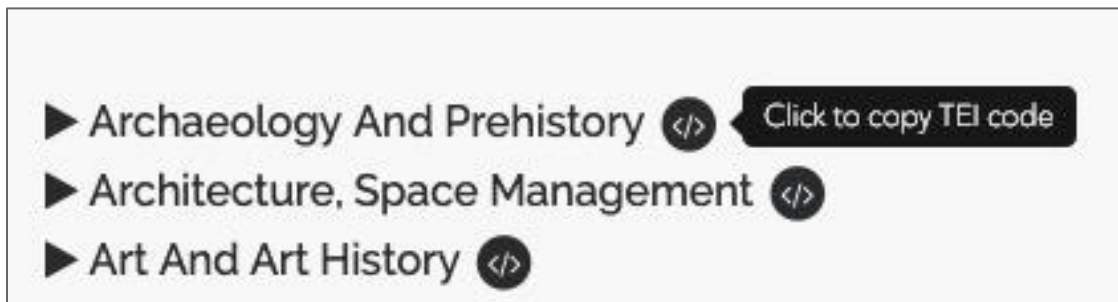
@type	activity, technique, object	standard	discipline
@source	tadirah	sisk	aurehal
@key See glossary	ex: Annotating Encoding Manuscript	ALTO-XML CMDI EAD	Communication sciences Geography Literature



Keywords in TEI

<term> element. Doc: https://ssk.readthedocs.io/en/latest/2_ssktei.html#term-element

New trick in [SSK Glossary](#): A button copies directly the TEI code



```
<term type="activity" source="tadirah"  
key="Encoding"/>
```



Resources in TEI

Inside <linkGrp> elements. Doc:

https://ssk.readthedocs.io/en/latest/2_ssktei.html#linkGrp-element

@type	generalResources	projectResources	projectResources
@source		CODATA	DTA
@corresp		http://www.codata.org	http://www.deutschestextarchiv.de

```
<linkGrp type="projectResources" source="DTA"
corresp="http://www.deutschestextarchiv.de"/>
```




Resources in TEI

<ref> element. Doc: https://ssk.readthedocs.io/en/latest/2_ssktei.html#ref-element

@type	spec	service	paper
@source	zotero	zotero	zotero
@target See Zotero	T7672NJ8	8BD6FDKR	DVKJRRVU

```
<ref type="spec" source="zotero" key="T7672NJ8"/>
```

What's next? (early 2019)

- Browsing bibliography
- Creating an **account** :
 - to manage **bookmarks**
 - to customize **scenarios** (by combining existing steps from SSK's research scenarios)
- **Contributing** directly on the interface:
 - creation
 - edition
 - customization
- Accessing a **multilingual** interface

Sustainability

- Open-source code and GPL licence.
- Underlying data described in TEI and hosted on GitHub under the licence CC-BY
- Bibliographical resources are part of a Zotero open library.
- DARIAH working groups: take over intellectual maintenance.



Test the SSK!



<http://ssk.huma-num.fr>



<http://ssk.readthedocs.io>



ssk@inria.fr