

Digital synapses: on by default

Magdalena Turska

TEI Publisher/e-editiones

ORCID 0000-0002-7060-3531

This paper proposal discusses how offering sensible defaults in a scholarly publication framework influences interoperability and collaboration in digital humanities projects from the bottom up.

DH research usually involves working with very diverse datasets and collections created in specific contexts, available in particular formats and stored in separate institutions. Let's take an editorial project as an example: from digital facsimiles, through the results of HTR or OCR analysis, to transcription, collation, named entity recognition and linking with external authority files to manual annotation and, eventually, publication online or in print. For each of these steps we are looking at a variety of tools, formats and protocols to choose from. In principle, this is a wonderful moment for humanities computing, where so much more is within our relatively easy reach than even just a decade ago and these opportunities are widely embraced in our community. On the other hand, the fundamental problem remains that creation of a digital resource is an intrinsically complex subject, as measured by a random assortment of acronyms from TEI to IIF, and the average humanities scholar can't be realistically expected to have a solid overview of the ever fluctuating technological landscape around us.

And yet she is still ultimately responsible for designing and orchestrating complex, collaborative workflows and has to directly face the short- and long-term consequences of each decision. Needless to say, opportunities to ease the burden and find local support that is tailored to specificities of research projects is a rare exception, by far not a rule. Therefore researchers are often left to navigate the stormy waters of a digital project in the making alone, without a clear map but just a set of tips and tricks from the collective brain of the community, from papers and conference presentations to leveraging informal connections, in order to find solutions that could be potentially reused or adopted.

It is interesting to see what happens when a community of practitioners instead of accepting the situation as is works together firmly prioritizing the development of a generic, systematic solution above the specific, without compromising on the latter. In this paper I would like first to briefly present the chain of events which led from the conception of the TEI Simple with the TEI Processing Model to the development of the TEI Publisher¹ framework and foundation of initiatives like e-editiones² and Sources Online. More importantly, I would like to demonstrate the visible impact these collaborations had on the landscape of digital scholarly editions in recent years.

The TEI Simple project, commenced in 2014, made an observation that TEI "has been able to achieve its range of use by adopting a descriptive rather than prescriptive approach, (...) and by eschewing any attempt to dictate how the digital texts should be rendered or exchanged. However, this flexibility has come at the cost of relatively limited success in interoperability³". Even a decade later it remains a popular rant to complain how overwhelming TEI can be as an encoding system, with a steep learning

¹ Meier, W., Turska, M. et al *TEI Publisher*, <https://teipublisher.com/>

² E-editiones: editions with future, an international non profit scholarly society <https://www.e-editiones.org/>

³ Rahtz, S. et al, *TEI Simple* <https://teic.github.io/TEI-Simple/>

curve for beginners. Therefore main objectives of TEI Simple became to provide a smaller, leaner customization of the full TEI vocabulary accompanied by "a notation (as an extension to TEI's ODD metalanguage) for specifying processing rules for TEI encoded texts⁴" as well as an implementation of default processing rules for TEI. This notation has become later known as the TEI Processing Model and incorporated into the TEI standard itself.

Very shortly after, in early 2015, the first implementation of the Processing Model has been built⁵, which has eventually formed the heart of the TEI Publisher. This application framework for scholarly publications embraced TEI Simple's motto of "power to the editor" and "standardize where you can, customize where you must" together with Pareto's 80/20 rule, aiming to build on a solid foundations of open source standards and to cover the majority of common requirements out of the box, while leaving the users freedom to opt-out of default solutions and tailor the publication to their particular needs. As TEI Publisher is now in version 9 obviously the feature set covered by default has gradually grown, and the same can be said about Publisher's user base, which allows us to make some interesting observations.

Even the earliest version of the TEI Publisher already came with an application generator and a collection of sample TEI encoded files and Processing Model ODDs tailored to showcase their particular features. This approach for quick prototyping where a new application is generated on a basis of an existing example has been a consistent trait already among the early adopters. At the same time a less obvious phenomenon caught our attention as numerous projects employed also the encoding strategies implemented in TEI Publisher samples. Particularly successful have been

⁴ *ibid.*

⁵ Meier, W. et al, *TEI Processing Model Toolbox* <https://github.com/wolfgangmm/tei-simple-pm>

the correspondence examples where both original transcription and aligned translation are available.

Arrival of faceted searches in TEI Publisher resulted in a wide adoption of this feature among its user community, the same can be said about popularity of facing transcription and facsimile display layout, and, more recently, adoption of the visual annotation editor as an integral part of the editorial workflow. Nevertheless it is important to note that a feature selection clearly remains a considerate and thoughtful process on behalf of the editors and developers. What we can see, is that some of the examples or features are very often incorporated into actual editions while others are considered more of a novelty and rather omitted in finished publications. Still, the availability of various examples, not only for distinct domains but also with varying encoding strategies, presentational tactics and styles, allows users to realize the possibilities and make an informed choice for their project.

We could see a strong influence with pretty much every new major feature and sometimes with the launch of specific editions as well. After Van Gogh Museum & Huygens ING released to the public the complete encoded corpus of the famous painter's correspondence, a demo app⁶ with an interactive document layout inspired by the original Van Gogh edition but implemented with the Lego-block, modular style of Publisher has found numerous followers since⁷. Despite Van Gogh's demo flamboyant appeal, the crown among textual *influencers* nevertheless goes to SSRQ⁸, a collection of Swiss

⁶ *Van Gogh Letters*, a TEI Publisher demo <https://teipublisher.com/exist/apps/vangogh>

⁷ Cf. e.g. *Electronic Edition of the works of Jean-Joseph Rabearivelo (french and malagasy)* <https://rabearivelo.huma-num.fr/exist/apps/jjr/index.html> or *Basler Edition der Bernoulli-Briefwechsel* (in preparation) https://bebb.jinntec.de/B_1750-12-29_991170430456905501

⁸ *Sammlung Schweizerischer Rechtsquellen* online, <https://editio.ssrq-online.ch/>

Law Sources Online, whose original application from 2018 has been effectively cloned wholesale by a number of similar projects in Switzerland⁹.

It was precisely the success story of SSRQ and, both local and international, collaborations which led to the foundation of e-editiones, a scholarly society dedicated to breaking the silos and bringing together editions, cultural heritage institutions and developers to pool resources and collaborate to empower our projects and sustain them far into the future. Born in the middle of the global pandemic, the society not only gathers its official institutional and individual members but includes much larger informal, grassroots community with 350 members of the lively Slack communication channel and more than 500 followers on social media, a regular monthly gathering online and, last but not least, a growing listing of Publisher-based projects. At the time of writing, about 40 projects chose to register, we are aware of a similar number still in preparation and estimate perhaps even another hundred that we never yet heard about.

As demonstrated above, the availability of an open source framework which allows users to mix and match existing solutions, strengthened by the existence of a supportive community of practice has proven very successful so far. The constant, informal dialogue through the Slack channel and monthly meetups, gives real forum to many diverse perspectives and allows our users to quickly identify issues, find assistance from others as well as partners for more formal collaborations sometimes leading to successful funding proposals. From a developer perspective, such an environment gives us a unique insight into the variety of use case scenarios for the framework we are building, therefore we can design more generic and more modular solutions which are form the solid foundation for specific projects. In this spirit we have the best of both worlds: standardized, professionally

⁹ Cf. e.g. *Die Urkunden und Akten des Klosters und des Oberamts Königsfelden* <https://www.koenigsfelden.uzh.ch/exist/apps/ssrq/index.html> or *Rechtsquellen des Kantons Zürich* <https://rechtsquellen.sources-online.org/index.html>

build components that can be orchestrated to create sustainable customized applications.

What we saved for last are the less obvious but equally important effects that TEI Publisher had on editions but also other digital humanities projects. As modularity and standardization are our two basic principles, we are building on standards and common libraries, therefore at the same time promoting them and benefitting from their availability: this applies to the TEI Processing Model itself, supported XML vocabularies like TEI, JATS or DocBook; IIIF for the images together with OpenSeaDragon¹⁰ and Tify¹¹ viewers; Verovio¹² for sheet music; MathML or TeX for mathematical formulas; OpenAPI¹³ for the specification of the programmatic interface and DTS¹⁴ for machine-consumption of digital text collections.

As a consequence, each and every TEI Publisher-based application exposes clearly specified and documented programmatic interface which can be usually accessed just suffixing the app name with `api.html`¹⁵. Similarly, each and every TEI Publisher app exposes endpoints for a DTS protocol and, from TEI Publisher 9 it also exposes endpoints for local authority registries. With all of this available as default and by design integrated into the framework we achieve interoperability out of the box, and treat it as a priority, not an afterthought for which there's never time.

¹⁰ OpenSeadragon: an open-source, web-based viewer for high-resolution zoomable images <https://openseadragon.github.io/>

¹¹ TIFY: a slim and mobile-friendly IIIF document viewer <https://tify.rocks/>

¹² Verovio is a fast, portable and lightweight open-source library for engraving Music Encoding Initiative (MEI) music scores into SVG. <https://www.verovio.org/index.xhtml>

¹³ Open API Specification <https://swagger.io/specification/>

¹⁴ The Distributed Text Services (DTS) Specification defines an API for working with collections of text as machine-actionable data. Publishers of digital text collections can use the DTS API to help them make their textual data Findable, Accessible, Interoperable and Reusable (FAIR). <https://distributed-text-services.github.io/specifications/>

¹⁵ Compare default TEI Publisher API and customized version from *Digitale Edition der Reisetagebücher von Johann Conrad Fischer 1794–1851* <https://www.johannconradfischer.com/api.html>

Custom API 1.0.0 OAS3

modules/custom-api.json

This is the place to add your own path endpoints

Servers

http://localhost:8080/exist/apps/rqzh2 - Endpoint for testing on localhost

Authorize

view

GET	/about/	Retrieve the HTML template used for displaying a document	vapi:view
GET	/people/{editionseinheit}/	Landing page for people	vapi:html
GET	/people/{editionseinheit}/{name}	Retrieve the HTML template used for displaying person metadata	vapi:html
GET	/organization/{editionseinheit}/	Retrieve HTML template: organisation-list.html	vapi:html
GET	/organization/{editionseinheit}/{name}	Retrieve HTML template: organization-detail.html	vapi:html
GET	/keyword/{editionseinheit}/	Landing page for keywords	vapi:html
GET	/keyword/{editionseinheit}/{name}	Retrieve the HTML template used for displaying keyword metadata	vapi:html
GET	/places/{editionseinheit}/	Landing page for places	vapi:html
GET	/places/{editionseinheit}/{name}	Retrieve the HTML template used for displaying places metadata	vapi:html

For a practical demonstration of the power but this approach affords, we'd like to present a project called Sources Online. It is an infrastructure for digital scholarly text editions based on TEI Publisher and IIF operated by the Trägerschaft Archives Online association in cooperation with e-editiones, Staatsarchiv Zürich, Karl Barth-Gesamtausgabe and Jinntec. In a nutshell, this is a search portal which can run queries against a number of independent editions at the same time and present aggregated results with advanced faceted and field-based search options. This is possible thanks to standardized API endpoints exposed by all the editions and demonstrates the interoperability potential for editorial projects despite their heterogeneous data and structure.

In the final illustration we can see the effects of searching for a word "Dorf", meaning "village" across all editions currently available in Sources Online. As shown below, there are matches in 3 of 5 editions: the correspondence of Alfred Escher, the Karl Barth edition, as well as the sources on the history of Zurich.

The screenshot shows the 'Sources Online' search interface. At the top left, the title 'Sources Online' is displayed. On the top right, there is a language dropdown set to 'English' and a 'Login' button. Below the title, a search bar contains the text 'dorf'. To the right of the search bar, there are navigation arrows and the text 'Found 67 items'. On the left side, there are several filter tabs: 'Fulltext', 'Title', 'Person', 'Place', 'Date: From', and 'Date: To'. The 'Fulltext' tab is active. The main search results area shows two entries, both starting with '1 Alfred Escher > Alfred Escher an Jakob Escher, Belvoir (Enge, Zürich), Dienstag / Samstag, 28. Mai 1839, 01. Juni 1839'. The first entry includes a link to 'dorf' and a snippet: '...ngen längs dem Venusberge nach Kessenich, Dottendorf, Fries (Blumer's und mein gewöhnlicher Reitweg) längs dem Rhei...'. The second entry includes a link to 'Dorf' and a snippet: '...isirt worden. Auf die Rothenthurm die Horner & besondersdie Landsgemeindezogen nun durch das Muottathaler mit H...'. On the right side, there are two sections: 'Edition' and 'Person'. The 'Edition' section has a 'Show top 50' checkbox and a list of items with counts: 'Alfred Escher-Briefedition' (14), 'Karl Barth Gesamtausgabe' (52), and 'Quellen zur Zürcher Geschichte' (1). The 'Person' section also has a 'Show top 50' checkbox and a list of items with counts: 'Escher (vom Glas) Alfred' (11), 'Escher (vom Glas) Jakob' (2), 'Heer Oswald' (1), 'Gerhardt, Paul (1607-1676)' (16), 'Goethe, Johann Wolfgang von (1749-1832)' (9), 'Barth, Karl (1886-1968)' (15), 'Luther, Martin (1483-1546)' (15), and 'Thurneysen, Eduard (1888-1974)' (13). At the bottom left, there are buttons for 'SEARCH' and 'NEW SEARCH'.

This simple example represent very complex paths through which it came into existence. From serendipity involved in bringing together all the people and institutions who made it possible to careful orchestration of tools and resources necessary to provide the technical scaffolding such a design requires. Openness is a key to it all: from open source libraries and datasets to collaborative attitude of our colleagues all the way to the open communication channels of programmatic interfaces that are always there to listen and respond to our requests. If we keep these synapses on by default, we're good to go for the long haul.